

Our research is inspired by the findings of the paper *TopicGPT: A Prompt-based Topic Modeling Framework*, which highlighted the limitations of the Mistral model in comparison to other approaches. We aim to build upon the foundational aspects of the Mistral model, addressing its shortcomings and enhancing its capabilities to improve performance.

Related works

In the domain of topic modeling and text categorization, a variety of methodologies have emerged over the years, each aiming to improve our ability to discern and categorize underlying themes within large datasets of text. Latent Dirichlet Allocation (LDA) [1], introduced in the early 2000s stands out as one of the foundational techniques that is still used in modern comparisons and benchmarks. LDA has paved the way for understanding topic distributions within documents through a generative statistical approach. However, the evolution of NLP technologies has ushered in a new era of topic modeling methods, characterized by enhanced sophistication and effectiveness.

In the paper *BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure* by Maarten Grootendorst, BERTopic is introduced as an advanced topic modeling technique. This method distinguishes itself by utilizing pre-trained transformer-based language models to generate and cluster document embeddings, in this way creating coherent topic representations through a novel class-based TF-IDF approach. Unlike conventional models such as LDA, BERTopic emphasizes the semantic relationships between words, ensuring more meaningful clustering and topic representation. The study highlights BERTopic's strengths in generating semantically coherent topics, its adaptability with different language models, and its capacity for dynamic topic modeling, showcasing its superior performance in various benchmarks.

BERTopic was used in the paper *ChatGPT in education: A discourse analysis of worries and concerns on social media* [2], which addresses the challenge of integrating ChatGPT in education, emphasizing the need to balance its educational benefits against potential risks and ethical concerns. It highlights areas such as academic integrity and the development of student skills. The researchers employed BERT-based topic modeling to analyze Twitter data, identifying key themes and concerns related to ChatGPT's application in education.

In *Top2Vec: Distributed Representations of Topics* by Dimo Angelov [3], Top2Vec is presented as an unsupervised model that innovatively uses document and word embeddings to discover topics within large text collections. It simplifies topic discovery by automatically determining the topic count and eliminating the need for stop-word removal, stemming, or lemmatization. The model's key strengths include its automated discovery of topics through semantic embeddings, creating a semantic space that accurately represents the similarities among documents, words, and topics. Top2Vec also introduces a topic evaluation metric based on mutual information, with which its ability to yield more informative and corpus-representative topics is shown. Additionally, it facilitates hierarchical topic reduction, aiding in the simplification of analysis without considerable information loss. Overall, Top2Vec marks a notable shift towards leveraging distributed representations for more efficient and intuitive topic modeling, indicating a promising avenue for further exploration in natural language processing and information retrieval fields.

A promising paper which offers a lot of space for improvement is TopicGPT [4], which represents a breakthrough in topic modeling. It utilizes a prompt-based framework with large language models (LLMs) to surpass traditional methods like LDA in terms of interpretability and semantic control. This method produces topics that are more aligned with human understanding and categorization. TopicGPT operates through a two-stage process: initially, it generates topics by prompting an LLM with a selection of documents and a list of previously generated topics, refining these topics to reduce redundancy and eliminate infrequent ones. In the second stage, the model assigns new documents to these topics, citing evidence from the documents to support the assignments and enhance the method's verifiability and interpretability. This approach is described as human-centric, focusing on creating intuitive topic structures and understandable document-topic associations. It also allows for user intervention in the generation process through seed topics and manual editing, ensuring the output closely aligns with user expectations.

The paper further discusses the use of Mistral, an open-source LLM, for topic assignment in an attempt to reduce dependence on more costly proprietary models such as GPT-3.5-turbo. Although Mistral shows reasonable effectiveness in topic assignment, it does not achieve the same level of performance in topic generation as its proprietary counterparts, highlighting the challenges of relying solely on open-source models for high-quality topic modeling.

There have also been a few papers trying to offer a comparative analysis of current state-of-the-art methods. In *Topic Modeling: A Consistent Framework for Comparative Studies* by Ana Amaro and Fernando Bacao [5], the authors tackle key challenges in topic modeling, presenting a comprehensive analysis aimed at enhancing the consistency and comparability of the algorithm evaluations of topic modeling. This work features a detailed comparative study of five TM algorithms over three benchmark datasets, evaluated against five distinct metrics. It updates the survey of approaches and metrics in topic modeling, introduces state-of-the-art algorithms like Top2Vec not covered in previous literature, and proposes a consistent framework for algorithm comparison. Key findings include Top2Vec's superior performance across all datasets when evaluated with Context Vectors (CV) Topic Coherence, suggesting newer approaches can provide more informative topics than traditional models such as LDA. The study emphasizes the importance of selecting appropriate evaluation metrics, pointing out the variability in algorithm performance under different conditions and highlighting the necessity of a comprehensive evaluation framework. The paper marks a significant contribution to the field by proposing a framework that ensures comparability and consistency in algorithm evaluation, revealing the potential of advanced embedding techniques in improving outcomes. It underscores the ongoing need for rigorous comparison and evaluation of algorithms to keep pace with emerging models and techniques.

Similarly, in the study *Leveraging State-of-the-Art Topic*

Modeling for News Impact Analysis on Financial Markets: A Comparative Study by Chen et al. [6], the authors employ a novel “News Impact Analysis” (NIA) framework, aiming to automate and streamline the process of evaluating news impacts on stock prices, addressing a gap in finance-specific news analysis literature. The paper scrutinizes three contemporary topic modeling methods — LDA, Top2Vec, and BERTopic — across a dataset comprising 38,240 financial news articles.

The findings indicate that BERTopic outperforms its counterparts by delivering higher coherence scores, improved interpretability, and reasonable computation times, with minimal data preprocessing requirements. The authors argue for the superiority of BERTopic due to its advanced embedding techniques and class-based TF-IDF procedure, which significantly enhance topic quality and relevance.

The work in the aforementioned papers provides us with the groundwork and the necessary tools to implement a solution to our problem, as well as evaluate the solution properly and compare it to existing approaches.

Corpus analysis

The primary dataset comprises information collected from an online discussion centered around the narrative “The Lady, or the Tiger” and is available on our Github repository¹. Each entry within the dataset includes the user’s pseudonym and their corresponding message. To ensure consistency, messages were categorized with a high level of inter-rater reliability. Coders were assigned the task of labeling entries with specific classifications, with the main classification being *Discussion type*. They were provided with definitions and examples of various discussion types to aid in this categorization process. Additionally, based on the identified discussion type, coders further marked messages with classifications such as *Dialogic spell*, *Uptake*, *Question*, and *Pivot*. For analysis, we used the *R1 Discussion Type* where it was available and replaced it with *R2 version* where it was not available. We can use the visualization depicted in Figure 1 to illustrate the popularity of each specific discussion type. Notably, the most prevalent type is *Seminar*, as defined in the provided codebook as *Discussion on the meaning or interpretation of content*. Following closely is *Deliberation*, with half the frequency of the former, defined as *Turns related to decision-making about the content*.

Additionally, we can analyze the *Pivot* column, which denotes messages that altered the course of discussion, with the formal definition being *The posting establishes or changes the procedural direction of the discussion*. This is evident in Figure 2, where the *Seminar* node occupies the central position with the highest number of connections, aligning with the prevalence of Seminar-type discussions. Outgoing connections are depicted with green arrows, indicating that most pivots originate from *Seminar* and transition into *Delibera-*

¹<https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-azb>

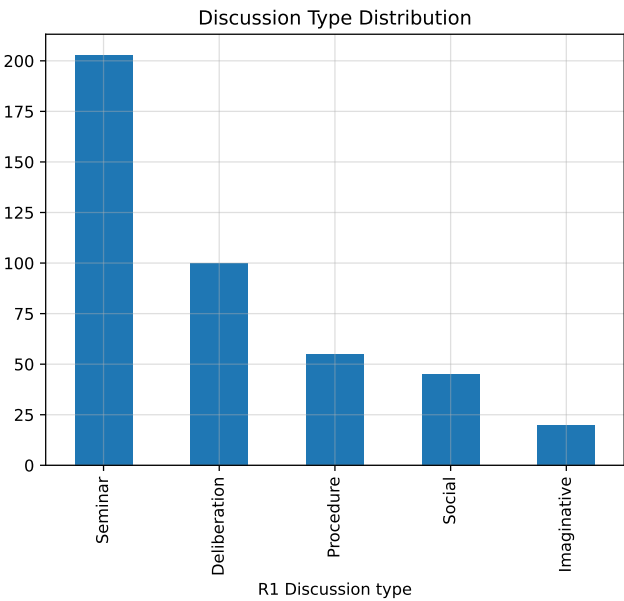


Figure 1. Distribution of different discussion type. We can predominantly see ‘Seminar’ type, which is expected from the definition and the type of discourse. Least common type is the “Imaginative” which “places the learner in the discussion as an active participant”.

tion discussions, and vice versa. This pattern aligns with the definition of *Deliberation* as *Turns related to decision-making about the content*, which often follows or precedes discussions about the story in the *Seminar* type.

Methods

Results

Discussion

Acknowledgments

References

- [1] D. Blei. Latent Dirichlet Allocation. 2003.
- [2] Lingyao Li, Zihui Ma, Lizhou Fan, Sanggyu Lee, Huizi Yu, and Libby Hemphill. ChatGPT in education: A discourse analysis of worries and concerns on social media. *Education and Information Technologies*, pages 1–34, 2023.
- [3] Dimo Angelov. Top2Vec: Distributed representations of topics, 2020.
- [4] Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework, 2023.
- [5] Ana Amaro and Fernando Bacao. Topic modeling: A consistent framework for comparative studies. *Emerging Science Journal*, 2024.

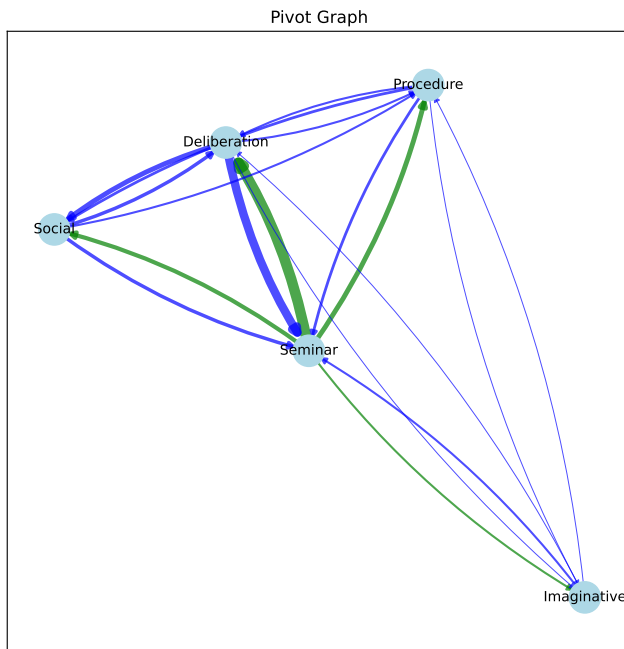


Figure 2. The graphical representation illustrates various pivot transitions, with nodes representing different discussion types and lines indicating transitions between them. The width of the lines corresponds to the frequency of specific transitions, while the direction indicates the sequence of occurrences. Notably, thick lines between *Deliberation* and *Seminar* highlight frequent transitions between these two discussion types.

- [6] Weisi Chen, Fethi Rabhi, Wenqi Liao, and Islam Al-Qudah. Leveraging state-of-the-art topic modeling for News Impact Analysis on financial markets: A comparative study. *Electronics*, 12:2605, 06 2023.