# Topic Modeling Across S&P 500 Earnings Calls

**Team 202** Mark Emmenegger (ME) | Talha Sumra (TS) | Sumit Singh (SS) |
Kenneth Thorburn (KT) | Nathaniel Beare (NB) | Gordon Garisch (GG)

## 1. Introduction

Every quarter, publicly listed companies in the US conduct earnings calls to discuss their financial results with investors, analysts, and the media. In this project, we are developing a tool to extract themes from these discussions to derive important topics relating to economic, financial, technological, environmental, and social domains. Our data consists of transcripts from the largest 500 US companies (S&P 500) over the past 10 years. We apply topic modelling, clustering and Large Language Model (LLM) based analysis to identify, structure, and interpret these key themes. Findings from the best-performing model are visualized interactively in a dashboard.

Our tool targets institutional and retail investors, market analysts, executives, and policy makers. It aims to significantly enhance users' market intelligence by providing a structured way to track key market-driving themes in S&P 500 earnings calls.

## 2. Problem Definition

Our objective is to extract, structure, and visualize key themes from an extensive corpus of U.S. corporate earnings call transcripts. These transcripts encompass the last decade of S&P 500 constituents, incorporating both the presentation and Q&A sections. This results in a dataset comprising over 20,000 earnings calls transcripts. Each transcript includes, on average, 80 statements made by executives and analysts, spans numerous pages, and contributes to a total data size exceeding 2 GB.

The first core problem is to apply and evaluate unsupervised topic modeling methods on this large-scale financial text corpus. We implement three approaches, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and BERTopic, and assess them using both quantitative metrics and human evaluations. Since topic models produce flat, unlabeled outputs, we apply hierarchical clustering to uncover structure and use a Large Language Model (LLM) to assign topic and cluster names.

The second core problem is to visualize these structured themes interactively. We develop two Tableau dashboards that integrate a D3.js-based nested bubble chart alongside other charts, such as bar charts and a time-series plot, to explore thematic evolution over time and across sectors.

## 3. Literature Survey

Research increasingly applies Natural Language Processing (NLP) to earnings call transcripts to predict financial outcomes (Jiang, 2021). Recent models that combine text and audio from earnings calls have shown improved performance in volatility forecasting (Qin & Yang, 2019; Yang et al., 2020). These approaches indicate that earnings calls can be useful for identifying market themes and trends.

Regarding the application of unsupervised NLP to financial texts, two primary tasks are clustering and topic modeling. Clustering assigns documents to a single category, while topic modeling identifies multiple topics within documents and quantifies their significance (Boyd-Graber & Mimno, 2017). Early topic modeling techniques include Latent Semantic Analysis (LSA), which uses singular value decomposition on the document-term matrix (Deerwester et al., 1990), and its probabilistic extension, Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999).

A significant advancement in topic modeling has been Latent Dirichlet Allocation (LDA), a Bayesian model that represents documents as distributions of topics and topics as distributions of words (Blei et al., 2003; Blei & Lafferty, 2009). Another widely used method is Nonnegative Matrix Factorization (NMF), a deterministic factorization technique known for consistent results across multiple runs (Choo et al., 2013; Kim & Park, 2011). More recently, BERTopic has become popular, utilizing transformer-based embeddings, dimensionality reduction, a clustering stage, and weighted tokenization of topics (Devlin et al., 2019; Grootendorst, 2022). While LDA, NMF, and BERTopic each have strengths and weaknesses, a study by Egger and Yu (2022) found BERTopic to perform best on Twitter data. However, it remains uncertain if this finding applies to longer texts such as earnings call transcripts.

This project uses the state-of-the-art topic models outlined above but expands on them in two ways. First, while topic models typically produce flat topic structures, we aim to create a hierarchical structure. Although Wu et al. (2024) propose a similar approach, their method lacks easy implementation in Python, so we opt for simpler hierarchical clustering methods. Second, topic models generate unlabeled topics due to their unsupervised nature. To address this, we use a Large Language Model (LLM) for topic labeling. While Mu et al. (2024) applied LLMs directly to documents for topic modelling, we find this approach less feasible for large text corpora. Instead, we will apply LLMs only for labelling topics (word distributions) that have been identified by other methods.

Topic modeling generally generates numerous topics, each defined by a set of words, requiring innovative visualization techniques for large unstructured data (Cao & Cui, 2016). Various practitioners have explored visualizing topic models, often using interactive elements (see Maskat et al., 2023 for an overview). Word clouds are one popular approach (Choi et al., 2018; Kherwa & Bansal, 2020). Other authors have emphasized circular packing diagrams and hierarchical clustering representations, notably Wang (2006), Chaney (2012), and more recently Kim (2020).

Our visualization layer integrates several of these proposed visuals in a Tableau Dashboard. In particular, the dashboard includes a D3.js-based visual that combines circle packing (Wang 2006) with interactive features for exploring topics and identifying themes in earnings transcripts.

## 4. Proposed method

**Data:** Historical S&P 500 constituents were sourced from the public sp500 GitHub repository and enriched with Company Name, GICS Sector, and GICS Sub-Industry data from the s-and-p-500-companies repository. This time-series data was then used to retrieve quarterly earnings call transcripts from the API Ninjas v1/earningscalltranscripts endpoint, covering Q1 2014 through Q4 2024.

**Topic Modelling:** As is typical for natural language processing, topic models require extensive data preprocessing.

The first preprocessing step, text cleaning, aims to filter out unnecessary words and sentences. This includes removing numbers, first names, short words (with some exceptions such as "AI"), punctuation, and short sentences. We also lemmatized our data. Lemmatizing converts words to their base forms, reducing the inflectional variability in texts. While these preprocessing steps are essential for NMF and LDA, BERTopic can theoretically work with raw text as it internalizes preprocessing by producing document embeddings with a transformer model. For the earnings call data, however, it turned out that BERTopic works better with cleaned textual data.

A second preprocessing step is feature extraction, which generates a numerical representation of the cleaned text for models to work with. This involved converting the cleaned text into a document-term matrix (DTM). While LDA requires DTMs with simple counts, NMF works better with DTMs that use a weighting scheme. Unlike LDA and NMF, BERTopic internalizes the feature engineering step. In the case of all three models, our feature engineering includes case-insensitive conversion (convert all text to lowercase), considers unigrams (single words) and bigrams (expressions consisting of two words), disregards stopwords (common words that typically do not influence meaning), and restricts the vocabulary to the 1,000 most frequent words.

After preprocessing, we ran our three topic models: scikit-learn's NMF and LDA, as well as BERTopic. While using the default version of NMF, which has only a few meaningful hyperparameters, we customized our Bayesian LDA model with priors commonly used in literature (Blei et al. 2003). Specifically, for our alpha prior that controls the document-topic distribution, we used 50/n (where n is the number of topics). For the beta prior that controls the topic-word distribution, we used 0.01. In the case of the BERTopic model, we had to use a highly customized version to ensure comparability to the other models. While we left some key building blocks like the transformer/embedding block unchanged, we specifically used k-means instead of HDBSCAN for the clustering block to better control the number of topics. Additionally, BERTopic in its default version assigns only one topic to each document, but we used the module's approximate_distribution() function to get the full document-topic distribution. This makes it analogous to the MNF and LDA models' output. All three models result in two matrices that we normalized for comparability: a document-topic matrix, which contains an importance distribution of topics for each document, and a topic-word matrix, which represents an importance distribution of words for each topic.

The primary hyperparameter for all three models is the number of topics. Consequently, we executed the models across a range of topic numbers deemed reasonable based on the corpus size and the intended use of the resulting topics. Specifically, we fine-tuned the models for 50, 100, 150, and 200 topics. Given the resource-intensive nature of preprocessing and topic model fitting, we conducted the tuning process on 10% of the dataset (2,000 earnings calls). These models were evaluated both quantitatively and qualitatively (see Section 5: Evaluation), and the best-performing model, Latent Dirichlet Allocation (LDA) with 100 topics, was subsequently applied to the full dataset (20,523 earnings calls). Due to limitations in memory and processing power, we executed the Python scripts in Azure ML Studio using the compute instance Standard_E4ds_v4 (4 cores, 32 GB RAM, 150 GB disk), which reduced the total runtime to a few hours.

**Clustering and Naming:** Unsupervised learning involves working with unlabeled data, which often produces outputs that are difficult to interpret. In our case, the topic modeling step generated topics that were both unnamed and unstructured. We addressed these challenges through an automated process, which is a key innovation of our project.

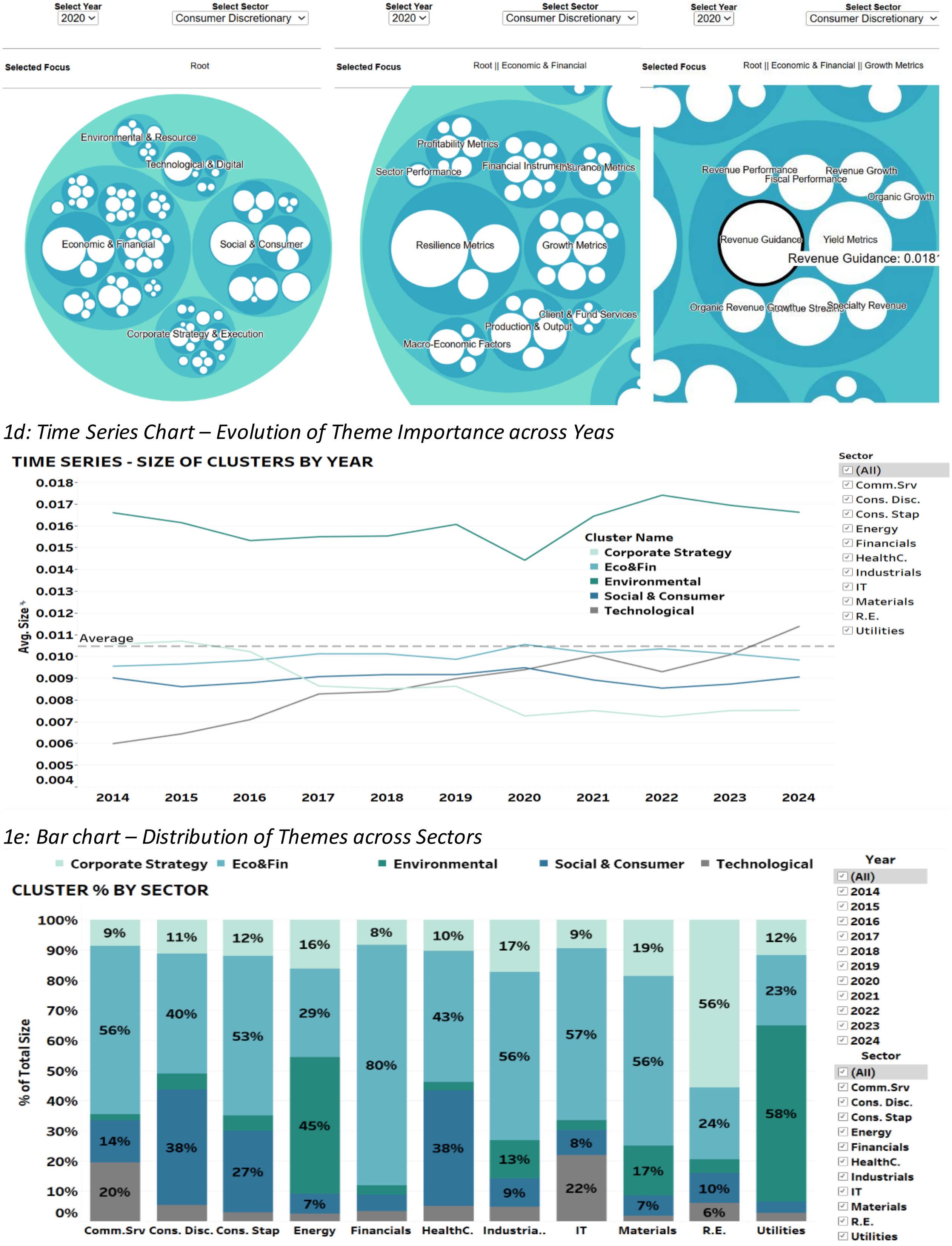
We began by naming the topics. Rather than manually labeling 100 topics, which is a standard but time-consuming task in supervised workflows, we leveraged GPT-4 via OpenAI's Python API. We automatically generated intuitive labels using carefully constructed prompts. Our pipeline first reduced each topic vector to its top N=5 most probable terms, then submitted these weighted lists to GPT-4, which returned concise (1–3 word) topic labels tailored to the context of earnings calls.

Subsequentially, we clustered the named topics into groups. We evaluated three methods: (1) *Baseline Clustering Approach:* We used scipy.cluster.hierarchy() to construct a dendrogram representing the hierarchical similarity between topics. From this, we derived the number of main clusters (level 1) and subclusters (level 2), grouping all 100 topics accordingly. Instead of assigning cluster names manually, we applied the same LLM-based method as used in topic labeling to name the clusters. (2) *End-to-End LLM Approach:* Here, we bypassed traditional clustering and used GPT-4 directly to create and name clusters. The model received the named topics along with their top N=5 terms and associated weights. (3) *LLM-Guided Hybrid Approach:* Building on the first method, we provided GPT-4 with the initial clustering results and tasked it with refining the groupings while naming the clusters, mirroring how a human analyst might iterate on the output of a classical clustering algorithm.

We experimented with varying cluster counts and prompts and ultimately selected the *LLM-Guided Hybrid Approach* for its coherence and interpretability. It resulted in a structured output of 5 main clusters and 29 subclusters across the 100 topics.

**Visualization:** Our visualization layer consists of two interactive dashboards built in Tableau, which show outputs from topic modeling, clustering, and LLM-based labeling (see Figure 1). The first dashboard explores cluster distribution across sectors and years. It includes a cluster percent by sector stacked bar chart, a top 10 topics table ranked by average size, and a time series chart tracking average cluster size evolution over the past decade. Global filters for sector and year allow dynamic slicing of the data. The second dashboard enables granular exploration, incorporating an interactive size-by-subcluster bar chart, and an embedded D3.js-based circle packing diagram.

**Figure 1: Selected Elements of the Tableau Dashboard**



*1a-c: Circle packing chart zoomed – a: Root Node, b: Main Cluster, c: Subcluster showing Topics (leaf nodes)*

*1d: Time Series Chart – Evolution of Theme Importance across Yeas*

*1e: Bar chart – Distribution of Themes across Sectors*

The D3.js-based circle packing diagram is a key innovation of our tool. This zoomable and filterable diagram employs the d3.pack() and d3.hierarchy() functions. The circles represent nodes in the hierarchy, with outer circles representing less granular nodes and inner circles representing more granular nodes, up to the leaf level. That is, given our clustering from above: The outer circle represents the 5 main topical clusters, the middle circle the sub-clusters, and the leaves the individual topics. The size of the circles indicates the relative importance of the topics. Topic importance is determined by the probability of a topic being contained in a document. When aggregating across documents (by company, sector, year, or quarter), the probability is averaged. The integration of the d3.js visualization in Tableau was achieved by hosting the HTML file on Github Pages and embedding the hosted webpage as an object in Tableau.

## 5. Evaluation

**Testbed**: Our project encompasses both a data science component (including topic modelling, clustering, and naming) and a tool component (an interactive dashboard). Consequently, we conducted a comprehensive set of experiments:

- Exp. 1: Evaluating the performance of various topic models on our data.
- Exp. 2: Assessing the meaningfulness of the topic names.
- Exp. 3: Determining the meaningfulness of the clustering.
- Exp. 4: Ensuring that the overall results (topic distributions) align with the ground truth.
- Exp. 5: Evaluating the usability and appearance of the tool (or dashboard).

We approached Experiment 1 by employing traditional data science methodologies, including fitting different candidate models for a range of topic numbers and evaluating them using both quantitative and qualitative measures. Experiments 2 to 5 were addressed through an online survey, conducted with subject matter experts in finance and economics analysis.

**Exp. 1: Evaluation of various topic models:** We tuned our three candidate models LDA, NMF and BERTopic by testing 50, 100, 150, and 200 topic numbers. We quantitatively evaluated these models by using the UMASS coherence and topic diversity: The UMASS coherence score compares the co-occurrence probability of word pairs in the same topic with the probability of these word pairs occurring by chance. The score usually ranges from negative values to zero; numbers closer to zero mean better coherence , i.e. more meaningful topics. Generally, scores around -1 or higher (closer to zero) are considered satisfactory. Topic diversity measures the ratio of unique words across all topics to the total number of words. Scores range from 0 to 1; values closer to 1 indicate that topics share fewer words (i.e. are more distinct and diverse). In addition, we qualitatively evaluated these models by performing intruder analysis. In this analysis, the top 5 most important words for each topic are selected and amended by one random word. A human then assesses whether they can spot the intruder word. If this is easy, the topics are well defined.

The result of this evaluation is displayed in Figure 2. The LDA model with 100 topics on the training data scores among the best in the quantitative evaluation: it belongs to the models with the highest UMASS coherence score (-0.99) and achieves still a reasonable topic diversity (0.64). It also scores best in our qualitative analysis. In addition, the LDA model is more interpretable than the other two candidates; Therefore, we decided to select LDA 100 as our model for the tool. Note that we reran the quantitative and qualitative tests for LDA 100 on the full data, which resulted in a slightly better performance than the model on the sample data (see also Figure 2).

Our result contrasts with the study mentioned in the literature survey by Egger and Yu (2022), where they find BERTopic to lead to better results on Twitter data. This difference could be due to two factors: First, earnings call transcripts are longer documents than tweets, making them suitable for LDA. Second, we aim at a large number of topics, while BERTopic's intermediate dimension reduction and clustering steps tend to induce sparsity.

**Exp. 2 to 5: Survey-Based Evaluation:** Experiments 2 through 5 were evaluated through an online survey conducted with 7 subject matter experts in finance and economics. The results are presented in

Figure 3. Experiment 2, which analyzed the fit of topic names to the defining keywords for a small sample of topics, received an average rating of 4.2 out of 5, indicating that the topic names were generally appropriate. Experiment 3 assessed the meaningfulness of the topic clusters and achieved the highest score of 3.9, which suggests that clusters are logical and cover the range of expected themes. For Experiment 4, which evaluated the alignment of topic distributions with the ground truth, the overall score was 4.0. A more detailed analysis showed lower ratings for alignment across time (3.3) and across sectors (3.6). This is likely due to increased variability in smaller subsamples. Finally, Experiment 5 rated the overall usability and appearance of the tool at 4.0. Among specific visual components, the circle packing chart was rated highest at 4.6, while the time series chart received the lowest score of 3.4.

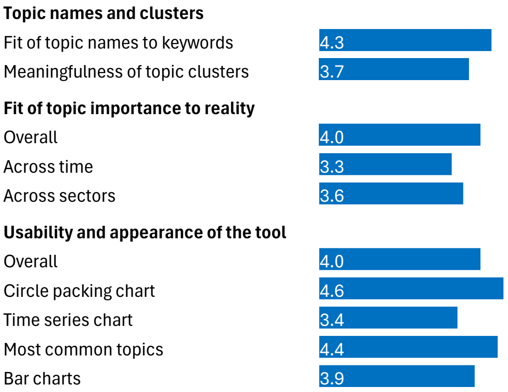
**Figure 2: Evaluation of Topic Modelling**

| **Model** | | | **Quant. measures** | | **Qualit. measure** |
|---|---|---|---|---|---|
| Type | No. Topics | Data | umass | topic diversity | Intruder analysis |
| LDA | 50 | sample | -1.12 | 0.84 | 4 |
| NMF | 50 | sample | -1.04 | 0.84 | 4 |
| BERTopic | 50 | sample | -1.24 | 0.96 | 3 |
| LDA | 100 | sample | -0.99 | 0.64 | 5 |
| LDA (*) | 100 | full | -0.98 | 0.66 | 5 |
| NMF | 100 | sample | -1.01 | 0.65 | 4 |
| BERTopic | 100 | sample | -1.22 | 0.86 | 3 |
| LDA | 150 | sample | -0.99 | 0.53 | 3 |
| NMF | 150 | sample | -0.99 | 0.58 | 3 |
| BERTopic | 150 | sample | -1.24 | 0.71 | 2 |
| LDA | 200 | sample | -0.99 | 0.46 | 3 |
| NMF | 200 | sample | -1.01 | 0.53 | 2 |
| BERTopic | 200 | sample | -1.20 | 0.58 | 2 |

*Notes: (*) Final model. Sample data = random sample of 2'000 earnings calls (approx. 10% of the data); used for tuning due to run time issues. Full data = full sample of 20'523 earnings calls. Intruder Analysis: Result rated from 1 (worst) to 5 (best) by the modelers.*

**Figure 3: Survey Results**

*Average of results: 1 = low to 5 = high*



*Note: Survey conducted with 7 subject matter experts, i.e. financial and economic analysts.*

## 6. Conclusions and Discussion

This project outlines an approach to extracting, structuring, and visualizing themes in over 20,000 S&P 500 earnings call transcripts using topic modeling techniques. Among the methods tested, Latent Dirichlet Allocation (LDA) with 100 topics, supplemented by Large Language Model (LLM)-based naming and hierarchical clustering, was found to be the most coherent and interpretable solution. The results show that unsupervised NLP methods, when adjusted and combined with LLMs, can produce a semantically sound structural analysis of earnings call transcripts. This offers a way to monitor key financial and strategic topics over time and across industries.

The developed dashboard translates these insights into an interactive interface, which received positive feedback in surveys with subject matter experts. This positions the tool as a potential asset for investors, analysts, and policymakers seeking to observe market themes in earnings calls systematically.

However, there are some limitations. Certain topics lack coherence and some subclusters may appear misaligned or overly narrow. Human oversight, especially from domain experts, could improve topic relevance through filtering, relabeling and regrouping. Additionally, the clustering methodology and its integration with LLM-based naming have room for further refinement.

Future research may investigate more sophisticated methods for integrating traditional clustering algorithms with LLMs. Additionally, comparing various LLMs in this context and employing a broader range of prompting strategies could yield valuable insights. To maintain relevance, operationalizing regular updates to the model pipeline will also be essential.

Please note that all team members have contributed a similar amount of effort.

## Bibliography

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In Text mining (pp. 101-124). Chapman and Hall/CRC.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3), 143-296.

Cao, N., & Cui, W. (2016). *Introduction to text visualization* (Vol. 1). Paris: Atlantis Press.

Chaney, A. &. (2012). Visualizing topic models. In *Proceedings of the International AAAIConference on Web and Social Media*, 6(1), 419-422.

Choi, M., Shin, S., Choi, J., Langevin, S., Bethune, C., Horne, P., ... & Choo, J. (2018). Topicontiles: Tile -based spatio-temporal event analytics via exclusive topic modeling on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-11).

Choo, J., Lee, C., Reddy, C. K., & Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12), 1992-2001.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7, 886498.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794.* (Note: This paper is not yet peer-reviewed and is not included in our count).

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57).

Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, 184, 115537.

Kherwa, P., & Bansal, P. (2020). Topic modeling: a comprehensive review. EAI Endorsed Trans. Scalable Inf. Syst., 7(24), e2.

Kim, J., & Park, H. (2011). Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6), 3261-3281.

Kim, H. D. (2020). Architext: Interactive hierarchical topic modeling. *IEEE transactions on visualization and computer graphics*, 3644-3655.

Maskat, R., Shaharudin, S. M., Witarsyah, D., & Mahdin, H. (2023). A Survey on Forms of Visualization and Tools Used in Topic Modelling. *JOIV: International Journal on Informatics Visualization*, 7(2), 517-526.

Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248.* (Note: This paper is not yet peer-reviewed and is not included in our count).

Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL 2019) (pp. 390–401).

Wang, W. W. (2006). Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 517-520.

Wu, X., Pan, F., Nguyen, T., Feng, Y., Liu, C., Nguyen, C. D., & Luu, A. T. (2024, March). On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 19261-19269).

Yang, L., Ng, T. L. J., Smyth, B., & Dong, R. (2020, April). Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020* (pp. 441-451).