# Unstructured Language Across Organizational Hierarchies: Evidence from the Enron Email Dataset

**Desiree Bengalli, Andrea Guarnieri, Tancredi Liani, Andrea Lisci, Giada Poloni**
Bocconi University

## Abstract

This study analyzes the Enron Email Dataset to investigate how internal communication patterns evolved across hierarchical levels as the company approached collapse. After preprocessing and cleaning the data, we begin with an exploratory data analysis that reveals distinct linguistic behaviors among low-, medium-, and high-level employees. Building on these differences, we train a supervised Support Vector Machine to predict the sender's hierarchical level based on linguistic features such as verbosity and lexical diversity, obtaining a 91%, 68% and 54% recall for high, medium and low senders respectively. However, given the difficulties in finding an optimal separation among unstructured data a Gradient Boosting model is implemented, better capturing non-linear relationships, obtaining 90%, 67% and 68% recall for high, medium and low level senders.

Following classification, we shift from stylistic to semantic analysis. We study sentiment evolution over time, using transformer-based models to track how emotional tone changes across levels, and link major fluctuations to external events in Enron's timeline. To further explore thematic structure, we apply Latent Dirichlet Allocation (LDA), but due to its limitations, we adopt BERTopic. Its results display improved coherence scores as measured by cosine similarity, and a clearer alignment with key moments in the company's downfall. Finally, to further assess cross-level differences, we analyze topic overlap among hierarchical groups, highlighting within the same organization key differences and similarities as the crisis approached.

This in-depth analysis helps illuminate the social dynamics of large organizations and how they evolve during crisis situations.

## 1 Introduction

Big companies can be described as complex systems, much like society itself. Rules, norms, and power dynamics lie at the heart of modern corporations, which play an increasingly central role in today's world. But how do these systems behave? How pronounced are the distinctions along their hierarchical scale? And how do they respond to internal and external shocks?

To explore these questions, we leverage one of the most famous corporate collapse of the early 2000s. Enron Corporation, a major electricity and energy provider, was one of the most prosperous U.S. companies in the 1990s. However, in 2001, the discovery of large-scale accounting fraud caused the company to go bankrupt.

Within the present analysis we deploy the Enron Email Corpus, first introduced by Klimt and Yang (2004), who preprocessed and released it in a structured format for applications in machine learning, text classification, and social network analysis. Since then, the dataset has become a benchmark for various investigations.

Our work builds on this body of research, with the aim of testing the following hypothesis: stylistic, emotional, and semantic layers jointly reveal how organizational hierarchy affects internal communication. To investigate this, we explore multiple layers of lexical styles and semantics to uncover how organizational hierarchy shapes internal discourse. By analyzing structural markers such as verbosity and lexical diversity, emotional tone via sentiment analysis, and thematic content through topic modeling, we aim to capture a multidimensional view of language use across roles.

## 2 Data Cleaning and Preprocessing

We begin by organizing the raw dataset, extracting key elements such as timestamp, sender, recipients, and the body of each email. One of the main challenges at this stage is identifying the individuals behind each email address, since names were often not explicitly stated. To address this, we focus on emails from trusted domains, such as Enron, Outlook, Yahoo, and Hotmail, and analyze the structure of the addresses to infer names wherever possible, looking for matches with a list of known Enron employees.

We then manually assign each sender to a hierarchical level, based on publicly available information. In total, we label 191 employees, with 19 executives, 75 middle managers, and 97 employees at the operational level.

We then clean the body of each email to remove irrelevant or noisy content, including replies from previous messages, automatic signatures, legal disclaimers, and system-generated text, as outlined by Tang et al. (2005).

## 3 Exploratory Data Analysis

After feature engineering and tokenization, we analyze several summary statistics and, as expected, when analyzing verbosity, we find that executives tend to write shorter and more concise messages.

During lexical analysis, we examine the most frequent words across the three levels, both before and after lemmatization. The vocabulary used at different levels reflects their organizational roles: top management tends to focus on coordination, crisis management, and formal communication, whereas lower-level employees use more operational language, centered around task execution and reporting. We also explore lexical diversity using two metrics: MTLD (Measure of Textual Lexical Diversity) and MATTR (Moving-Average Type-Token Ratio). The results do not show substantial differences across groups, suggesting that while content and tone differ by role, the breadth of vocabulary remains relatively consistent across the whole organization.

Named Entity Recognition reveals that several institutional and regulatory actors frequently appear. Entities such as the Federal Energy Regulatory Commission (FERC), which investigates Enron's trading practices, and the Public Utilities Commission (PUC), central in discussions on price manipulation in California, are among the most cited (for a more detailed representation, please refer to Figure 1)

## 4 Classification among Levels

Building on the evidence emerged in the EDA phase we now aim to verify the current hypothesis: language and communication habits reflects organizational role. To verify this, we implement two supervised classifiers: a Support Vector Machine (SVM) and a Gradient Boosting model (XGBoost). Each email serves as the unit of classification to infer the sender's hierarchical level.

In both models we include a set of features such as TF-IDF vectors to capture lexical content, whether the email was sent between 8 p.m. and 6 a.m. (as a proxy for executive work habits), number of recipients, number of emails addressed to executives, verbosity dummies, email length, average word length, and presence of courteous or directive keywords (e.g., "please," "should," "crisis," "order"). To validate the performance we split the data into into training (67%) and test (33%) sets.

### 4.1 Support Vector Machine

The first implementation relies on a Support Vector Machine. We select this model because it performs well on high-dimensional, sparse input data, such as our email corpus. The model achieves 91% recall on high-level emails, 67% on medium, and 54% on low. Although effective at detecting top-level messages, its performance drops sharply on lower levels. This is due to the fact that those at the operational level are more numerous and heterogeneous, therefore more lexically diverse and less clearly separable.

### 4.2 Gradient Boosting

To improve performance, we adopt Gradient Boosting (XGBoost), a tree-based ensemble method that builds decision trees sequentially. This model handles non-linear interactions and heterogeneous feature distributions more effectively. As expected, Gradient Boosting outperforms SVM, with classification recall of 90% for high, 67% for medium, and 68% for low (for a more detailed comparison of the results, please refer to Table 1). Feature importance scores reveal that email length, recipient structure, and time of day are among the most predictive variables (for a more detailed representation, please refer to Figure 2).

The particularly high performance in classifying top-level (high) emails is likely due to the distinctiveness of executive communication. High-level emails are fewer in number and tend to follow consistent linguistic and structural patterns (e.g. shorter and more directive phrasing) and, as a result, the model finds it easier to identify them.

## 5 Sentiment Analysis

After analyzing lexical patterns and stylistic differences across hierarchical levels, we shift our focus to content-related distinctions. Rather than examining how employees write, we now explore their tone and the prevailing sentiment within their communications.

Previous research has applied sentiment analysis to the Enron email dataset. However, these studies, such as Belay et al. (2024), have not identified clear patterns, likely due to aggregating emails across roles with divergent communication functions. Our analysis builds on these results by distinguishing between hierarchical levels with the aim of uncovering more precise dynamics.

We apply two transformer-based models, RoBERTa and DistilBERT (both configured to produce a score between -1 and 1), to compute sentiment for each individual email and then aggregate scores on a bimonthly basis to observe temporal trends by hierarchical level.

Although the absolute values differ, the models exhibit similar trends, including noticeable drops in sentiment around key crisis periods. More in detail, RoBERTa shows a marked increase in sentiment among high-level employees between January and March 2000, possibly reflecting optimism linked to the launch of Enron's high-speed broadband network or strong stock performance (Thomas, 2002). Conversely, we observe a sharp sentiment drop among low-level employees between March and May 2001. This may correspond to the collapse of the Blockbuster deal [Thomas (2002)]. Beyond such local shifts, both models also reveal a gradual decline in executive sentiment beginning in mid-2001, in line with the unfolding crisis. (for a more detailed representation, please refer to Figure 4 and Figure 3)

## 6 Topic Modeling

Diving deeper into content analysis, we now explore what employees write about, to identify recurring subjects of discussion and assess how the content varies between organizational levels.

### 6.1 LDA

We train separate LDA topic models for each hierarchical level: low, medium, and high. To identify the best configuration, we perform a grid search varying three parameters: number of topics ($k$), document-topic prior ($\alpha$), and topic-word prior ($\eta$). We explore values of $k \in \{4, 5, 6, 7, 8\}$, and test both `symmetric` and `asymmetric` settings for $\alpha$, as well as `symmetric` and `auto` for $\eta$. The optimal configuration is selected based on a coherence score which evaluates how frequently the top words in each topic co-occur across documents. More in detail, the best results are obtained with 7 topics for low-level employees, 5 for medium-level, and 4 for high-level, with both $\alpha$ and $\eta$ set to `symmetric` in all cases. This suggests that enforcing uniform priors over documents and words leads to more stable and interpretable topics in our setting. Furthermore, to evaluate topic effectiveness, we compute pairwise cosine similarity between the c-TF-IDF representations of each topic's most relevant words, excluding self-comparisons. This metric acts as a proxy for consistency, with higher values indicating tighter lexical grouping. While these scores are reasonably high, the resulting dominant topics mostly revolve around meetings and workflow automation, suggesting limited semantic depth and little substantive insight overall.

### 6.2 BERTopic

Models like LDA rely on word co-occurrence patterns, which often fail to capture meaningful structure in short, unstructured, and informal texts such as emails and ignore context and semantic similarity. In contrast, transformer-based models generate contextual embeddings that better reflect the meaning of text segments, even when documents are brief or stylistically diverse. Therefore, to analyze topics more robustly, we adopt BERTopic. To generate the embeddings, we rely on BAAI-llm, a SentenceTransformer model optimized for semantic similarity and well-suited for handling short, informal texts like emails.

However, even with these adjustments, using HDBSCAN proves unstable in our case. The structure of the Enron emails does not produce well-separated clusters, resulting in fragmented and inconsistent topic assignments. Therefore, to improve topic stability, we replace it with KMeans, a centroid-based algorithm that allows for more deterministic clustering which requires the number of clusters (topics) to be specified in advance. To effectively determine this, we use the same grid search procedure previously applied to LDA and, for each hierarchical level, we evaluate models with 8, 10, 12, and 15 topics, selecting the configuration that maximizes average coherence.

The results show that the optimal number of topics varies by level: 15 for low, 8 for medium, and 12 for high. This pattern reflects the structure and communication needs of each group. Low-level employees tend to engage in a wider range of operational tasks, which results in greater thematic dispersion. In contrast, medium-level managers often act as intermediaries, communicating within more narrowly defined functional areas. Executives, while fewer in number, deal with a broader set of strategic and institutional issues, which likely explains the number of topics observed in the high group.

To assess the semantic coherence of each topic generated by BERTopic, we once again compute intra-topic cosine similarity. The results show a slight improvement compared to those obtained with LDA, indicating more effective grouping (for a more detailed comparison of the result please refer to Table 2).

To further validate the quality of the extracted topics, we analyze their evolution over time. We compute the temporal trend of each topic and visualize its prevalence across the months leading up to the crisis. The three graphs show considerable topic variance, each marked by distinct spikes. For low-level employees, a peak in the energy variance reporting topic emerges around May–June 2001, possibly reflecting broader national concerns, including a decline in renewable energy use (Wald, 2002) and the energy crisis in the U.S. Northwest that began in late 2000 (Northwest Power and Conservation Council, 2001). Among medium and high-level employees, spikes in electricity market and energy market instability topics both occur around January 2001. These movements likely relate to the 2000 California energy crisis (Federal Energy Regulatory Commission, 2005), when surging demand led major providers, including Enron, to restrict supply, triggering price hikes and intervention by the California Independent System Operator. Overall, peaks in topic activity correspond to known events in Enron's trajectory, such as regulatory investigations, market disruptions, reinforcing the effectiveness of the model in capturing not only semantic structure but also the dynamic nature of internal communication during the company's decline (for a more detailed representation, please refer to Figure 5, Figure 6 and Figure 7).

### 6.3 Topic Overlapping

To examine how content varies across hierarchical levels, we compare topic representations using c-TF-IDF. We compute cosine similarity between aligned topic vectors, applying a dual threshold using both Cosine and Jaccard similarity on the top-10 keywords. This method reveals no significant overlapping topics between levels, indicating that each group focuses on distinct content areas, a separation also clearly visible in the heatmaps and UMAP projection (please refer to Figure 8, Figure 9, Figure 10 and Figure 11).

These findings suggest that hierarchical position influences not only how people communicate, but also what they communicate about. Even during key shared events, such as the incoming crisis in the second half of 2001, employees at each level address the situation in distinct ways, reinforcing the idea that role shapes not only perspective, but also the language through which events are interpreted and discussed.

## 7 Conclusions and Related Works

Our initial hypothesis that communication content is shaped by one's role in the organization was confirmed both linguistically and semantically. These findings shed light on why previous analyses, such as earlier sentiment aggregation by Belay et al. (2024), may have yielded inconclusive or contradictory results. Ignoring the structural hierarchy means collapsing heterogeneous communication patterns into a single signal, averaging out the true meaning.

This theory finds support also outside the discipline of Natural Language Processing. For example, within the field of Network Analysis, Diesner et al. (2005), in their study of Enron's graph, reveal that employees occupy structurally distinct positions. More in detail, executives form tightly reciprocal clusters, while lower-level staff engage in sparse, hierarchical, and mostly upward directed exchanges. In other words, the very architecture of Enron's communication network mirrors its hierarchical divisions. Our results align with this structure: different ranks don't just speak differently, they speak about different things and failing to account for these discrepancies does not just weaken analysis, it risks misreading the organization entirely.

## References

Belay, A. et al. (2024). Sentiment analysis of enron emails: Challenges in role-agnostic aggregation. *Journal of Applied NLP Research*, 9(1):55–70.

Diesner, J., Frantz, T. L., and Carley, K. M. (2005). Communication networks from the enron email corpus: It's always about the people. Enron is no exception. *Computational & Mathematical Organization Theory*, 11(3):201–228.

Federal Energy Regulatory Commission (2005). The western energy crisis, the enron bankruptcy, and ferc's response. Technical report, Federal Energy Regulatory Commission.

Klimt, B. and Yang, Y. (2004). Introducing the enron corpus. In *CEAS 2004 – Conference on Email and Anti-Spam*.

Northwest Power and Conservation Council (2001). Energy crisis of 2000/2001. Web report.

Tang, J., Li, H., Cao, Y., and Tang, Z. (2005). Email data cleaning. In *Proceedings of KDD 2005*. ACM.

Thomas, C. W. (2002). The rise and fall of enron. *Journal of Accountancy*.

Wald, M. L. (2002). U.s. use of renewable energy took a big fall in 2001. *The New York Times*.

# A Appendix

Table 1: Performance Comparison: SVM vs Gradient Boosting

| Class | SVM | | | Gradient Boosting | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Low | 0.78 | 0.54 | 0.64 | 0.79 | 0.68 | 0.73 |
| Medium | 0.57 | 0.68 | 0.62 | 0.65 | 0.67 | 0.66 |
| High | 0.60 | 0.91 | 0.72 | 0.68 | 0.91 | 0.77 |
| Accuracy | | 0.65 | | | 0.72 | |
| Macro avg | 0.65 | 0.71 | 0.66 | 0.71 | 0.75 | 0.72 |
| Weighted avg | 0.68 | 0.65 | 0.65 | 0.72 | 0.72 | 0.71 |

Table 2: Topic Modeling Results: LDA vs BERTopic Across Hierarchical Levels

| **Level** | **LDA** | | **BERTopic** | |
|---|---|---|---|---|
| | **Label** | **Coh.** | **Label** | **Coh.** |
| Low | System Operations & Database Errors | 0.281 | Status Confirmations & Internal Feedback | 0.948 |
| | Schedule & Workflow Automation | 0.260 | Energy Variance Reporting | 0.928 |
| Medium | Transaction Processing & Credit Documentation | 0.498 | Casual Chats & Informal Humor Threads | 0.438 |
| | Daily Team Coordination | 0.437 | Enron Transactions & Counterparty Exposure | 0.436 |
| High | Contract Negotiations & Team Coordination | 0.513 | Unstructured Scheduling Amid Executive Chaos | 0.460 |
| | Executive Meetings & Communications | 0.493 | Last-Minute Invitations & Scrambled Executive Panels | 0.448 |

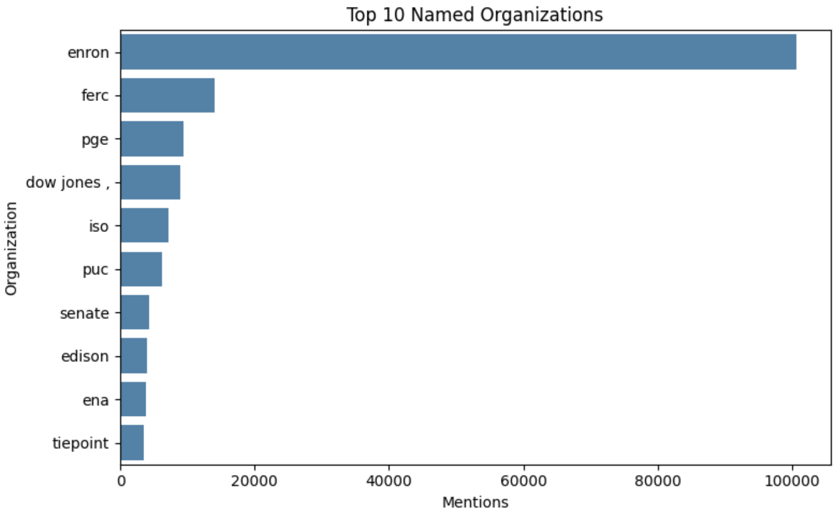


Figure 1: Named Entity Recognition

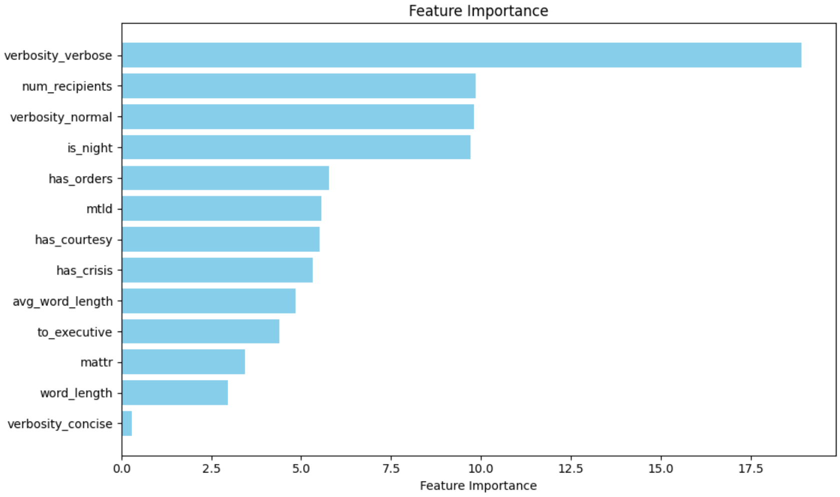


Figure 2: Feature importance (XGBoost)

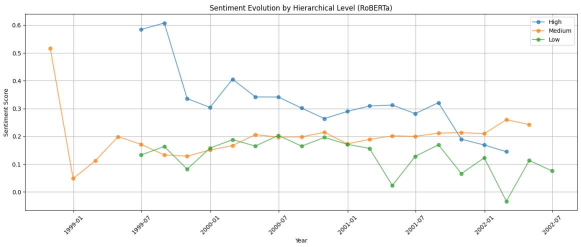


Figure 3: Sentiment evolution with RoBERTa

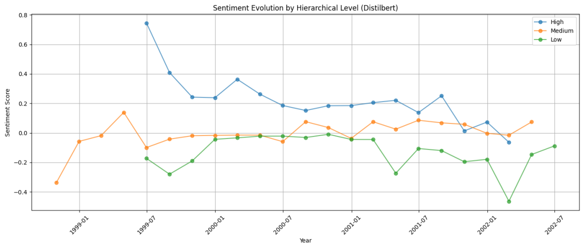


Figure 4: Sentiment evolution with DistilBert

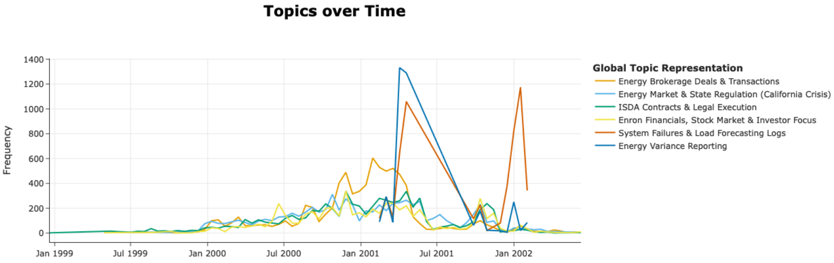


Figure 5: Topics over time for low level employees

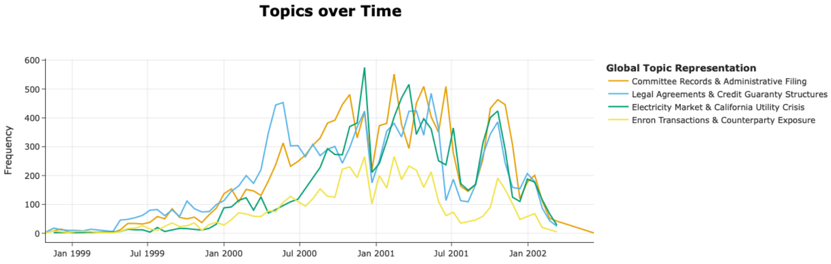


Figure 6: Topics over time for medium level employees

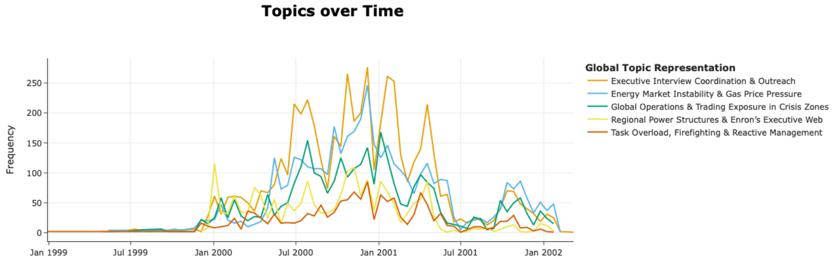


Figure 7: Topics over time for high level employees

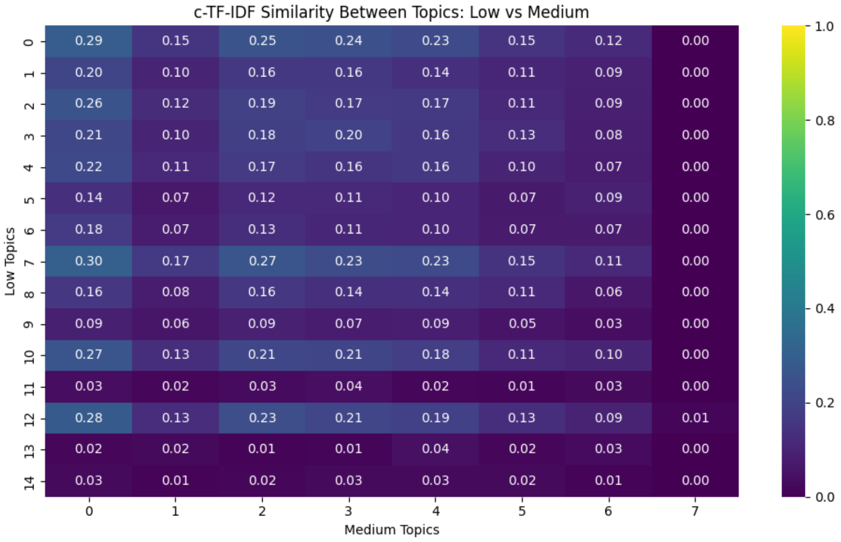


Figure 8: Similarity between topics: Low vs Medium

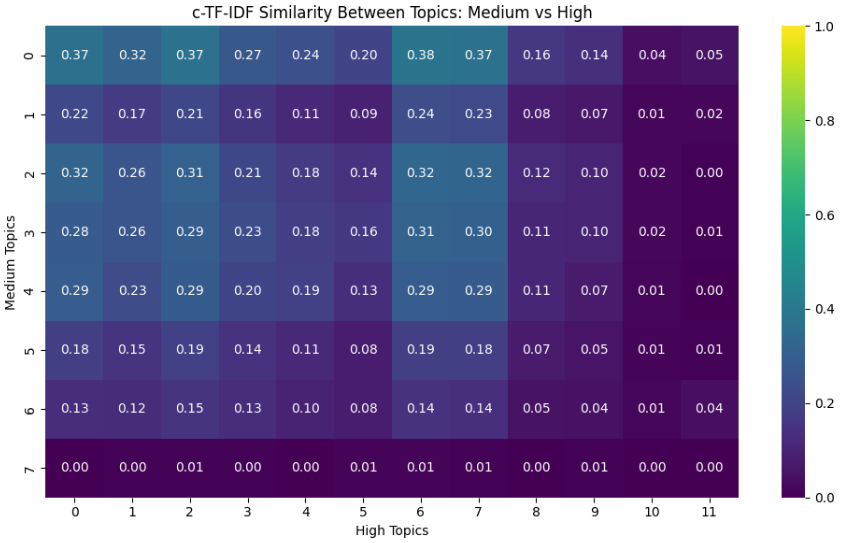
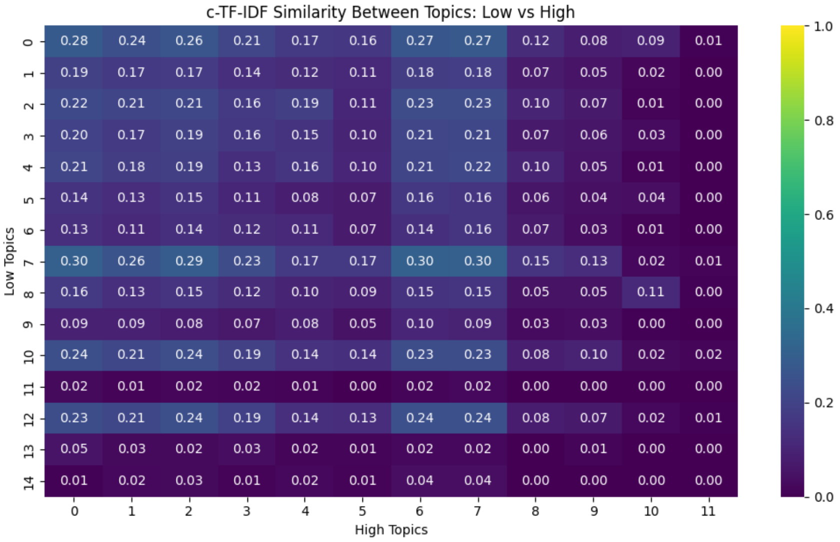


Figure 9: Similarity between topics: Medium vs High



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.28 | 0.24 | 0.26 | 0.21 | 0.17 | 0.16 | 0.27 | 0.27 | 0.12 | 0.08 | 0.09 | 0.01 |
| 1 | 0.19 | 0.17 | 0.17 | 0.14 | 0.12 | 0.11 | 0.18 | 0.18 | 0.07 | 0.05 | 0.02 | 0.00 |
| 2 | 0.22 | 0.21 | 0.21 | 0.16 | 0.19 | 0.11 | 0.23 | 0.23 | 0.10 | 0.07 | 0.01 | 0.00 |
| 3 | 0.20 | 0.17 | 0.19 | 0.16 | 0.15 | 0.10 | 0.21 | 0.21 | 0.07 | 0.06 | 0.03 | 0.00 |
| 4 | 0.21 | 0.18 | 0.19 | 0.13 | 0.16 | 0.10 | 0.21 | 0.22 | 0.10 | 0.05 | 0.01 | 0.00 |
| 5 | 0.14 | 0.13 | 0.15 | 0.11 | 0.08 | 0.07 | 0.16 | 0.16 | 0.06 | 0.04 | 0.04 | 0.00 |
| 6 | 0.13 | 0.11 | 0.14 | 0.12 | 0.11 | 0.07 | 0.14 | 0.16 | 0.07 | 0.03 | 0.01 | 0.00 |
| 7 | 0.30 | 0.26 | 0.29 | 0.23 | 0.17 | 0.17 | 0.30 | 0.30 | 0.15 | 0.13 | 0.02 | 0.01 |
| 8 | 0.16 | 0.13 | 0.15 | 0.12 | 0.10 | 0.09 | 0.15 | 0.15 | 0.05 | 0.05 | 0.11 | 0.00 |
| 9 | 0.09 | 0.09 | 0.08 | 0.07 | 0.08 | 0.05 | 0.10 | 0.09 | 0.03 | 0.03 | 0.00 | 0.00 |
| 10 | 0.24 | 0.21 | 0.24 | 0.19 | 0.14 | 0.14 | 0.23 | 0.23 | 0.08 | 0.10 | 0.02 | 0.02 |
| 11 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.00 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| 12 | 0.23 | 0.21 | 0.24 | 0.19 | 0.14 | 0.13 | 0.24 | 0.24 | 0.08 | 0.07 | 0.02 | 0.01 |
| 13 | 0.05 | 0.03 | 0.02 | 0.03 | 0.02 | 0.01 | 0.02 | 0.02 | 0.00 | 0.01 | 0.00 | 0.00 |
| 14 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.01 | 0.04 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 |

Figure 10: Similarity between topics: Low vs High

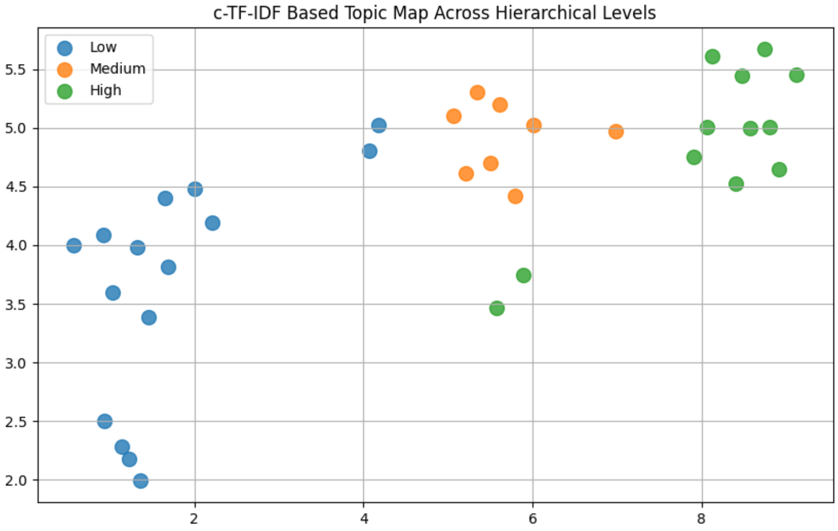


Figure 11: Topic Map with c-TF-IDF for levels