Write me project - Information Retrieval

Alessandro Barone

February 8, 2025

1 Introduction

The extraction of information from unstructured text represents a fundamental challenge in text mining, particularly when dealing with emails and chat messages. These communication forms present unique challenges due to their semi-structured nature, combining metadata, fixed formulas, and dynamic dialogical structures.

While emails and chats are ubiquitous in both personal and professional settings, their unstructured nature complicates automated information extraction. The meaning of messages often depends not only on their content but also on their context within broader conversations, requiring sophisticated analysis techniques. Recent literature has made significant progress in this field: from Agrawal et al.'s (2008) scalable entity extraction methods to Al-Moslmi et al.'s (2020) comprehensive overview of named entity techniques for knowledge graphs, and Hong et al.'s (2023) exploration of knowledge-grounded dialogue modeling. However, the specific challenges of email and chat analysis still require tailored approaches that can handle their unique characteristics.

2 Research Question and Methodology

This project aims to develop a comprehensive methodology for extracting information from email and chat messages, focusing on two key levels: non-structural elements (arguments and co.), and dialogical patterns (interaction dynamics between participants). Structural elements were also explored during the project, mainly to extract and divide the datasets, but they are not the main focus of the work.

The approach combines multiple NLP techniques including topic modeling (LDA and BERTopic), named entity recognition (using spaCy and BERT models), sentiment analysis (BERT) and dialogue analysis using different approaches. The methodology was evaluated on four distinct datasets: Clinton emails, Enron emails, fraudulent emails, and NPS chat . Each analysis technique required specific preprocessing approaches: topic recognition focused on text normalization and vocabulary construction, NER preserved case information and entity indicators, while sentiment analysis maintained emotional markers while removing technical content. All code and experimental results are available in a public GitHub repository, ensuring reproducibility and facilitating further research in this domain.

3 Experimental Results

The experimental analysis in this project is based on four primary datasets, each representing a distinct type of text data: emails and chat messages. Below is a brief description of each dataset:

- **Enron Email Dataset**: A collection of approximately 500,000 corporate emails from Enron Corporation employees, obtained during the Federal Energy Regulatory Commission's investigation.
- **NPS Chat Dataset**: Chat logs from the Naval Postgraduate School, containing multi-participant conversations with timestamps and user identifiers.
- **Fraudulent Emails Dataset**: Over 2,500 "Nigerian" fraud letters (419 scams) from 1998-2007, characterized by deceptive content aimed at financial fraud.
- **Hillary Clinton Email Dataset**: Approximately 7,000 pages of emails from Clinton's tenure as Secretary of State, released by the State Department and containing high-level professional communication.

3.1 Topic Recognition Analysis

For topic recognition, i experimented two distinct approaches: Latent Dirichlet Allocation (LDA), a traditional probabilistic model that views documents as mixtures of topics, and BERTopic, a more recent approach that leverages BERT's contextual embeddings combined with clustering techniques. While LDA relies on word co-occurrence patterns to identify topics, BERTopic first creates document embeddings using transformer models and then applies dimensionality reduction and clustering to identify topics, potentially offering more semantically meaningful results at the cost of higher computational requirements.

Metric	**Clinton**		**Enron**		**Fraud**		**NPS**	
	LDA	**BERT**	**LDA**	**BERT**	**LDA**	**BERT**	**LDA**	**BERT**
Total Documents	6,520	6,520	9,994	9,994	2,943	2,943	683	683
Number of Topics	6	10	6	10	6	10	6	10
Topic Diversity	0.85	-	0.83	-	0.73	-	1.00	-
Exec. Time (s)	8.32	26.32	10.49	81.53	5.86	27.98	1.76	6.80
Outliers (%)	-	31.2%	-	54.0%	-	24.0%	-	1.8%

Table 1: Comparison of LDA and BERTopic Performance Across Datasets

The comparison between LDA and BERTopic reveals distinct characteristics of each approach. LDA, being a probabilistic model, assigns every document to a topic mixture and shows good topic diversity scores, particularly high in informal communications (NPS chat). Running BERT, I choose to try searching more topics, to identify more deeper meanings. The outlier ratio is notably high in corporate emails (Enron: 54.0%) and lower in more structured communications (NPS chat: 1.8%). While BERTopic requires more computational resources, as shown by longer execution times, it provides more nuanced topic representations. LDA's seems in the end to be the most suitable choice for the datasets.

Metric	Clinton	Enron	Fraud	NPS	Mean
Coherence Score	0.454	0.412	0.389	0.536	0.448
Perplexity	5186.03	4823.15	3245.67	1064.38	3579.81
Topic Diversity	0.900	0.925	0.850	0.975	0.913
Execution Time (s)	4.857	3.245	2.156	0.150	2.602
Model Parameters (identical across datasets)					
Number of Components: 6					
Maximum Document Frequency: 0.950					
Minimum Document Frequency: 2					
Maximum Iterations: 15					

Table 2: LDA Model Performance Metrics Across Datasets

Table 2 presents the performance metrics and configuration parameters of the LDA implementation across all datasets. The coherence score measures the semantic interpretability of the discovered topics, with values closer to 1 indicating more coherent topics. Perplexity quantifies how well the model generalizes to unseen documents, with lower values suggesting better predictive performance. Topic diversity represents the uniqueness of discovered topics, measured as the proportion of unique words in the top N words across all topics. The execution time reflects the computational efficiency of the model training process.

Focusing on LDA, the datasets revealed distinct topical patterns across different types of communication. Here is presented a detailed analysis of the topics identified in each dataset

3.2 Clinton Email Dataset

The Clinton emails reveal distinct topics centered around political and diplomatic communications. The most prominent themes include international security relations (with references to Israel and women's issues), high-level political discourse (featuring terms like ”Obama,” ”Clinton,” ”president”), electoral politics (”party,” ”election,” ”percent”), and operational communications (”fyi,” ”print”). This mix of topics effectively captures both the strategic and day-to-day aspects of diplomatic correspondence.

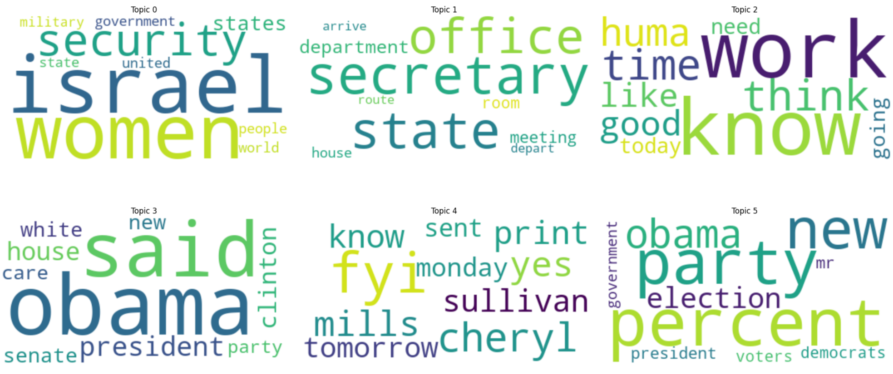


Figure 1: Clinton email dataset LDA visualization

3.3 Enron Email Dataset

The Enron corpus exhibits strong themes related to energy sector operations. Key topics revolve around energy pricing and contracts (with terms like "gas," "price," "deals"), energy management ("power," "electricity"), and business strategy ("trading," "company"). These topics clearly reflect Enron's core business activities and corporate communications structure.

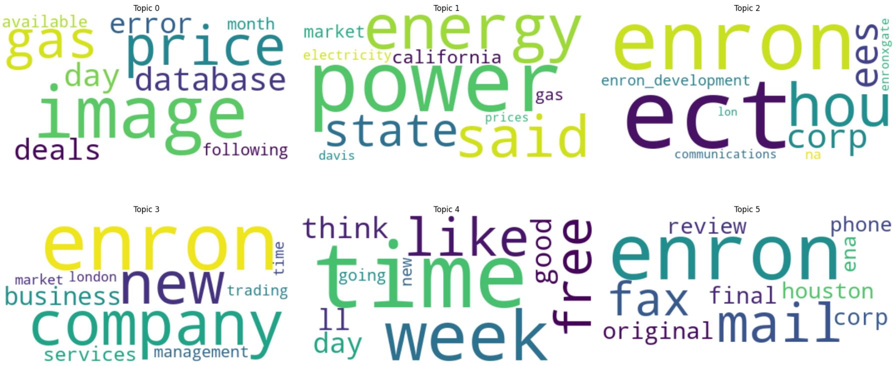


Figure 2: Enron email dataset LDA visualization

3.4 Fraudulent Email Dataset

Fraudulent emails show consistent patterns focused on financial deception. The topics cluster around financial transactions ("account," "transfer"), inheritance scenarios ("money," "bank," "kin"), personal narratives ("father," "company"), and urgent financial appeals ("funds," "late"). These patterns align with typical "419 scam" characteristics.

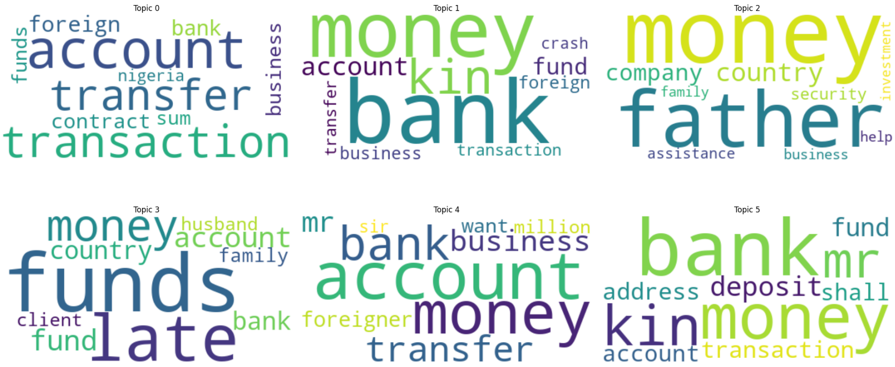


Figure 3: Fraudulent email dataset LDA visualization

3.5 NPS Chat Dataset

The NPS chat analysis reveals topics characteristic of informal online communication. The discourse includes social interactions ("join," "tell"), casual expressions ("hiya," "lol"), greetings, and temporal coordination ("time," "week"). While the informal nature of chat communications makes topic identification more challenging, the analysis effectively captures the spontaneous and social nature of online chat interactions.

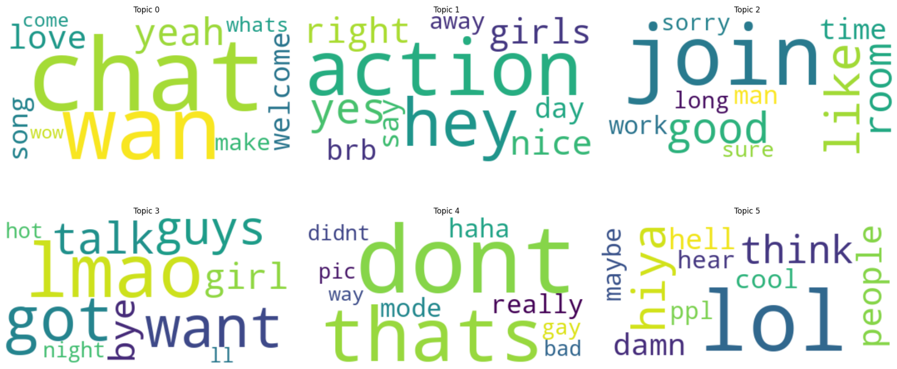


Figure 4: NPS Chat dataset LDA visualization

3.6 Named Entity Recognition Analysis

I implemented a dual-model approach to do NER:

1. **SpaCy Model**: Utilized the large English language model from spaCy, which excels at identifying standard entity types such as persons, organizations, locations, and dates. This model is particularly effective for well-structured professional communications.

2. **BERT-based Model**: Employed the BERT-large-cased model fine-tuned on CoNLL-2003, specifically the 'dbmdz/bert-large-cased-finetuned-conll03-english' variant. This transformer-based approach provides enhanced context awareness and better handling of ambiguous entities.

Both models were applied to process texts across the four datasets, The analysis was designed to identify several key entity types like: PERSON (Individual names and references), ORG (Organizations and institutions), GPE (Geopolitical entities), DATE (Temporal references), MONEY(Monetary values and references), LOC (Non-GPE locations) and so on.

Metric	**Clinton**		**Enron**		**Fraud**		**NPS**	
	BERT	**spaCy**	**BERT**	**spaCy**	**BERT**	**spaCy**	**BERT**	**spaCy**
Total Docs.	6,711	6,711	9,976	9,976	2,942	2,942	684	684
Total Entities	3,838	40,591	156,697	324,657	52,747	78,363	172	88
Entities/Doc	0.57	6.05	15.71	32.54	17.93	26.64	0.25	0.13
Entity Types	4	18	5	18	6	18	4	9
Exec. Time (s)	224.68	73.60	739.46	389.77	288.59	182.61	19.30	1.63

Table 3: Comparison of BERT and spaCy NER Performance Across Datasets

The comparison reveals significant differences between the two NER approaches. SpaCy consistently identifies more entity types (18) and generally detects more entities per document, likely due to its broader entity type classification system which includes temporal (DATE, TIME) and numerical (CARDINAL, ORDINAL) entities. BERT, while more conservative in entity recognition, shows consistent performance across datasets with higher execution times. The stark difference in entity detection is particularly notable in the Clinton emails (0.57 vs 6.05 entities/doc) and chat messages (0.25 vs 0.13

entities/doc), suggesting different sensitivity levels to formal versus informal text. The execution time difference favors spaCy, especially for larger datasets, making it more suitable for real-time applications.

This dual-model approach allows us to cross-validate entity detection and leverage the complementary strengths of traditional and transformer-based NLP architectures. The results provide insights into how different types of entities are used across various communication contexts.

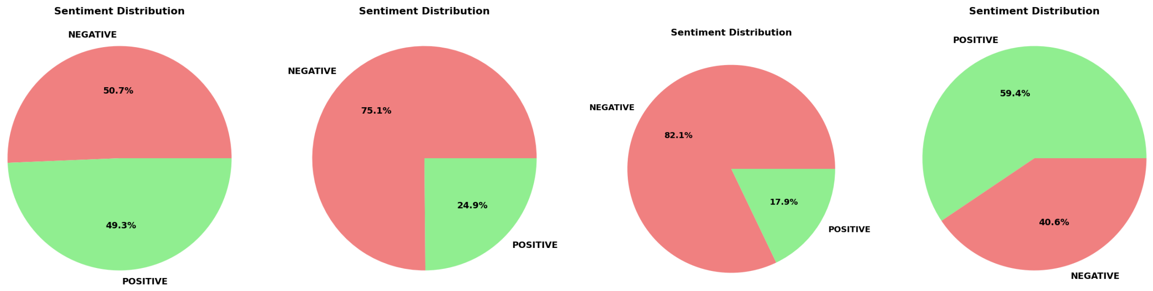
3.7 Sentiment Analysis

Sentiment analysis was performed to understand the emotional tone and polarity of communications across the datasets. I employed the DistilBERT model fine-tuned on the Stanford Sentiment Treebank v2 (SST-2) dataset, which classifies text into positive and negative sentiments while providing confidence scores for each classification.

The analysis pipeline processes texts in batches to optimize computational efficiency, with a maximum sequence length of 512 tokens. For each text segment, the model outputs:

- A binary sentiment label (POSITIVE/NEGATIVE)
- A confidence score ranging from 0 to 1

Figure 5a presents three visualizations of the sentiment analysis results for the Clinton email dataset:



(a) Clinton (b) Enron (c) Fraudulent (d) NPS chat

Figure 5: Sentiment Analysis across datasets

Metric	**Clinton**	**Enron**	**Fraud**	**NPS**
Total Messages	6,671	4,906	2,942	10,296
Sentiment Distribution				
Positive	3,292 (49.3%)	1,220 (24.9%)	526 (17.9%)	6,118 (59.4%)
Negative	3,379 (50.7%)	3,686 (75.1%)	2,416 (82.1%)	4,178 (40.6%)
Confidence Scores				
Pos. Mean	0.941	0.928	0.868	0.962
Neg. Mean	0.944	0.967	0.960	0.943
Pos. Median	0.992	0.985	0.934	0.996
Neg. Median	0.987	0.995	0.991	0.978

Table 4: Sentiment Analysis Results Across Datasets

The sentiment distribution reveals distinct patterns across the datasets. The Clinton emails show a nearly balanced distribution (49.3% positive, 50.7% negative), suggesting a neutral diplomatic communication style. In contrast, both the Enron (75.1% negative) and fraudulent emails (82.1% negative) exhibit strong negative sentiment biases, possibly reflecting the corporate crisis context and deceptive nature of these communications respectively. The NPS chat dataset shows a more positive tendency (59.4% positive), characteristic of casual social interactions. Notably, the model shows high confidence across all classifications, with median scores consistently above 0.93, indicating strong sentiment signals in the texts.

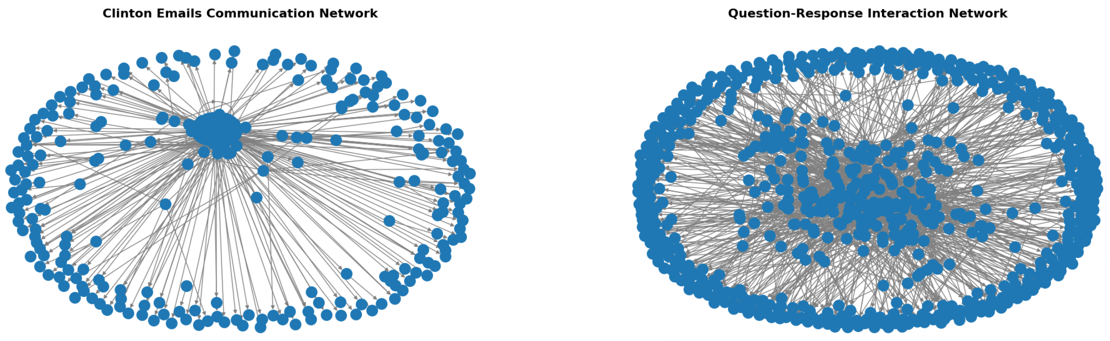
3.8 Interaction and Community Dialogue Analysis

My approach to analyzing communication patterns and community structures varied across datasets, adapting to their different characteristics and structures.

For the **Clinton, Enron**, and **fraudulent email** datasets, i implemented this approach to get interactions:

- **Node Definition**: Each unique email address or user identifier represents a node in the graph
- **Edge Creation**: Directed edges are created from sender to recipient(s)
- **Weight Calculation**: Edge weights are computed based on the frequency of interactions between pairs of users

For the NPS chat dataset, which lacks explicit sender-recipient relationships, i developed an alternative approach based on message similarity (using embeddings transoferms) and concatenating user information with the message content, questions, and temporal proximity. Here we can see Clinton emails and NPS chat interactions comparison, plotted using a directed graph with NetworkX:

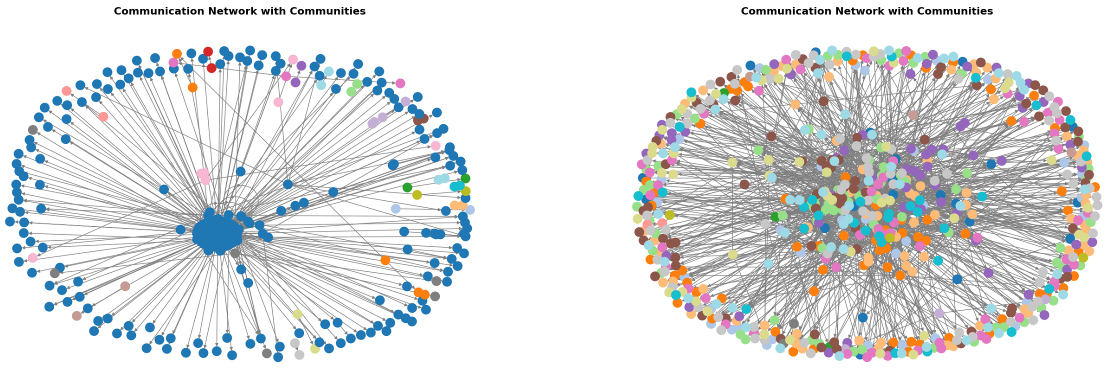

Clinton Emails Communication Network
Question-Response Interaction Network

(a) Clinton

(b) NPS

Figure 6: Interactions across datasets

For community detection in the interaction networks, i employed the Louvain algorithm, a hierarchical clustering method that optimizes modularity. This algorithm iteratively maximizes the density of edges inside communities compared to edges between communities. The method works in two phases: first, it optimizes modularity locally by moving nodes between communities, then it aggregates nodes of the same community to build a new network of communities. These steps are repeated iteratively until maximum modularity is achieved. In the next figure we can see how communities are identified between Clinton and NPS dataset:


Communication Network with Communities
Communication Network with Communities

(a) Clinton

(b) NPS

Figure 7: Communities across datasets

4 Concluding Remarks

The analysis of email and chat communications using various NLP techniques has revealed several interesting patterns. The multi-level approach to information extraction proved effective, with performance varying based on message formality and structure.

Named Entity Recognition showed significant variations across datasets, with spaCy excelling in entity type range (18 vs 4-6) and BERT showing higher confidence scores in formal communications (0.90 in Enron vs 0.75 in chat messages). Topic modeling comparison between LDA and BERTopic revealed complementary strengths: LDA provided complete coverage with good diversity scores (0.73-1.00), while BERTopic offered more nuanced identification despite higher outlier ratios (54% in Enron emails).

For dialogue analysis, my approach adapted to dataset characteristics. In email datasets, interaction networks were built using sender-recipient relationships, while for chat data, we leveraged message similarity and temporal proximity. The Louvain community detection algorithm effectively identified communication clusters, revealing distinct patterns between formal email exchanges and informal chat interactions.

Sentiment analysis highlighted characteristic patterns: Clinton emails showed balanced sentiment distribution, while fraudulent emails exhibited strong negative bias (82.1%). The high confidence scores (median 0.93%) suggest robust classification.

Future work could explore:

1. Integration of multiple NLP techniques for robust classification systems
2. Development of specialized preprocessing pipelines
3. Extension of dialogue analysis tracking dialogue state
4. Exploration of cross-lingual capabilities

The methodology provides a foundation for these developments, allowing easy integration of new techniques and improvements.