NLP - P8 - Write Me

Stefano Nava 27436A

Abstract

The extraction of meaningful information from unstructured email text is a critical challenge in the field of natural language processing (NLP). This project explores a two-step approach to the email content analysis: structural classification and topic modeling. First, emails were segmented into structural components (greetings, body, and closings) using pattern-based matching and regular expressions. This classification enabled a more granular analysis of the content and improved the consistency of the subsequent processing. After structural classification, the email body was cleaned through a preprocessing pipeline that included tokenization, stopword removal, and lemmatization. The text data was then vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) and processed with Non-Negative Matrix Factorization (NMF) to uncover distinct topics. The identified topics were evaluated based on their coherence and distribution within the dataset. The results demonstrate the ability of NMF to effectively cluster emails into meaningful categories, highlighting key patterns and recurring subjects. This combined approach provides a foundation for improving automated classification and content analysis of large volumes of email data, with potential applications in spam detection, email filtering, and content-based recommendation systems.

1 Introduction

The rapid increase in the volume of email communication has made it essential to develop efficient methods for extracting meaningful information from unstructured text. Emails are inherently complex due to their partially structured nature, which includes both formal elements (such as greetings, signatures, and metadata) and dynamic content that reflects the dialogical nature of communication. The ability to automatically analyze and categorize email content has significant applications in information retrieval, email filtering, spam detection, and business intelligence.

Natural Language Processing (NLP) provides powerful tools for analyzing unstructured text, but emails pose specific challenges due to their mixed structure and informal language. Previous research on email classification has focused on either structural analysis or semantic content extraction. Structural classification aims to segment emails into standardized components such as greetings, body, and closings, which enables better consistency in processing and improves the accuracy of the subsequent analysis. On the other hand, topic modeling allows for the discovery of hidden patterns and recurring themes within large datasets of textual content, revealing meaningful insights into the subject matter of the emails.

This project addresses both aspects by combining structural classification and topic modeling into a unified processing pipeline. The structural classification component segments each email into greetings, body, and closing parts using regular expressions and pattern-based matching. This improves the consistency and interpretability of the data for the next stage. The topic modeling component applies Non-Negative Matrix Factorization (NMF) to identify latent themes within the email bodies after a preprocessing phase involving tokenization, stopword removal, and lemmatization. This combined approach leverages the advantages of both structural analysis and semantic modeling to enable more effective and interpretable analysis of large-scale email datasets.

2 Research question and methodology

2.1 Research Question

The primary goal of this project is to develop an automated system for analyzing unstructured email text by addressing two fundamental challenges: (1) identifying the structural components of an email, and (2) discovering latent topics within the email body to uncover patterns and recurring themes. The research aims to answer the following questions:

- Can structural classification improve the consistency and quality of subsequent topic modeling on email content?
- How effectively can Non-Negative Matrix Factorization (NMF) identify meaningful topics from a dataset of email bodies after structural segmentation and preprocessing?

By addressing these questions, the project seeks to explore the potential of combining structural analysis and semantic modeling to improve the understanding and automatic processing of email data.

2.2 Methodology

The proposed approach is structured into two main phases: structural classification and topic modeling. Their combination is designed to improve the interpretability and performance of email analysis, providing a structured and semantically rich understanding of the dataset.

2.2.1 Structural classification

The first phase involves segmenting each email into three distinct structural components:

- **Greeting**: common opening phrases (e.g., ”Dear John,” ”Hi,” ”Hello”) are detected using regular expressions designed to match common patterns in business and personal email communication.
- **Body**: the main content of the email is extracted after removing the greeting and before detecting the closing section.
- **Closing**: common closing phrases (e.g., ”Best regards,” ”Sincerely,” ”Thank you”) are identified using regular expressions.

This segmentation enables more consistent preprocessing and helps reduce noise in the subsequent topic modeling phase.

2.2.2 Body preprocessing

After structural classification, the body of each email is processed through a text-cleaning pipeline that includes:

- **Tokenization** – Splitting the text into individual words.
- **Stopword Removal** – Removing common, non-informative words using a predefined list of stopwords from the nltk library.
- **Lemmatization** – Converting words to their base form using WordNetLemmatizer to reduce variability and improve clustering accuracy.
- **Character Cleaning** – Removing special characters and retaining only alphanumeric content and periods.

2.2.3 Vectorization and Dimensionality Reduction

The processed email bodies are vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to create a numerical representation of the text. TF-IDF reflects the importance of each term relative to the entire dataset.

To reduce the dimensionality of the data and improve the efficiency of topic modeling, Truncated Singular Value Decomposition (TruncatedSVD) is applied, retaining the most informative components of the vectorized data.

2.2.4 Topic Modeling

After dimensionality reduction, Non-Negative Matrix Factorization (NMF) is used to extract latent themes from the data. NMF is particularly suitable for text analysis because it imposes non-negativity constraints, which leads to a more interpretable and human-readable representation of topics.

The number of topics is defined through empirical tuning, and the most representative terms for each topic are extracted to provide a clear understanding of the discovered themes.

2.2.5 Evaluation

The evaluation of the topic modeling process is based on three key metrics: *Coherence Score*, *Topic Diversity*, and *Reconstruction Error*. These metrics provide a comprehensive assessment of the quality and consistency of the discovered topics, as well as the model's ability to represent the input data accurately. The evaluation is performed during the topic modeling phase to exploit the available data directly within the processing pipeline.

Coherence Score

The coherence score measures the semantic similarity among the terms within each topic. A high coherence score indicates that the top terms in a topic are semantically related, reflecting a well-defined theme. In this project, the coherence score is calculated using *cosine similarity* between the vector representations of the top n terms in each topic. The average similarity across all topics is computed as the final coherence score:

$$C = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i<j} \mathrm{sim}(w_i, w_j)}{\binom{n}{2}}$$

where:

- T = number of topics
- w_i, w_j = vector representations of terms within a topic
- $\mathrm{sim}(w_i, w_j)$ = cosine similarity between terms

A coherence score close to 1 indicates that the terms within a topic are highly similar, improving the interpretability of the results.

Topic Diversity

Topic diversity measures the degree of uniqueness of the terms among different topics. A high topic diversity score indicates that the topics are well-separated and cover different areas of the text. It is computed as the ratio between the number of unique terms in the top n terms across all topics and the total number of terms $(n \cdot T)$ in the top n list for all topics:

$$D = \frac{\text{number of unique terms}}{n \cdot T}$$

where:

- T = number of topics
- n = number of top terms per topic

A diversity score close to 1 means that the discovered topics are distinct and non-overlapping.

Reconstruction Error

The reconstruction error measures how well the model approximates the original data. It is computed directly from the NMF model as the Frobenius norm between the original input matrix X and the reconstructed matrix $\hat{X} = W \cdot H$:

$$E = ||X - W \cdot H||_F$$

where:

- W = document-to-topic matrix
- H = topic-to-term matrix

A low reconstruction error indicates that the model accurately represents the input data, while a high error suggests that the model is struggling to approximate the underlying structure of the text.

Evaluation Strategy

The combination of coherence, diversity, and reconstruction error provides a balanced evaluation strategy:

- A **high coherence score** reflects that the topics are internally consistent.
- A **high diversity score** indicates that the topics are well-separated and distinct.
- A **low reconstruction error** demonstrates that the model captures the key patterns in the data effectively.

In this project, the goal was to optimize coherence and diversity while maintaining a reasonably low reconstruction error. Fine-tuning the number of topics and the regularization parameters for NMF was critical to balancing these objectives.

3 Experimental results

This section provides an overview of the dataset used for the experiments, the evaluation metrics applied, and the results obtained. The goal of the experimental analysis is to assess the quality of the discovered topics and the consistency of the topic modeling approach.

3.1 Dataset Overview

The experiments were conducted on a dataset of **3332** emails related to fraudulent activities. The dataset includes a variety of scam formats, including advance-fee fraud (commonly known as ”Nigerian Prince” scams), fake lottery winnings, inheritance scams, and impersonation of government or bank officials.

Each email was preprocessed to remove metadata, stopwords, and non-alphabetic characters, followed by lemmatization to reduce morphological variation. The processed dataset was then transformed into a Bag-of-Words representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This preprocessing step was critical to reduce noise and improve the model’s ability to identify meaningful patterns in the data.

Dataset Property	Value
Total number of emails	3332
Average email length (tokens)	226.12
Number of unique terms after preprocessing	38233
Number of top terms per topic	10

Table 1 Summary of the dataset properties.

3.2 Model Configuration

The topic modeling was performed using Non-Negative Matrix Factorization (NMF) with the following configuration:

- Number of topics: **5**
- Maximum number of iterations: **1000**
- Number of top terms per topic: **10**

To evaluate the model performance, the following metrics were computed:

- **Coherence Score**: measures the internal consistency of the terms within each topic.
- **Topic Diversity**: evaluates the distinctiveness of the terms across different topics.
- **Reconstruction Error**: measures how well the model reconstructs the original dataset from the topic distribution.

3.3 Evaluation Results

Table 2 summarizes the evaluation results obtained from the experiments.

Metric	Value
Coherence Score	0.6395
Topic Diversity	0.90
Reconstruction Error	50.2019

Table 2 Summary of the evaluation results.

The obtained coherence score indicates that the terms within each topic are semantically quite similar, suggesting that the discovered topics are (partially) internally consistent. The topic diversity score reflects a high level of separation among the topics, meaning that the model was able to capture distinct patterns in the data. However, the reconstruction error remains relatively high, indicating that there is still room for improvement in the model's ability to approximate the original data distribution.

3.4 Topic Examples

The following table presents the top 10 terms for each of the discovered topics, reflecting the most frequent and meaningful terms extracted from the dataset:

Topic	Top Terms
1	ministry, nigeria, payment, foreign, transaction, fund, bank, transfer, contract, account
2	son, assistance, family, late, investment, company, money, security, country, father
3	plane, department, customer, crash, unclaimed, relation, money, deceased, bank, kin
4	shall, deposit, client, provide, branch, document, mr, attorney, bank, kin
5	know, dont, life, lord, money, organization, charity, want, god, husband

Table 3 Top 10 terms for each discovered topic.

The extracted topics reflect the most common patterns found in fraudulent emails:

- Topic 1 highlights impersonation of government and diplomatic officials ("ministry", "nigeria").
- Topic 2 identifies inheritance and financial scams ("son", "family", "late", "father").
- Topic 3 identifies accident-related unclaimed capitals ("plane", "crash", "deceased")
- Topic 4 identifies legally related matters ("deposit", "document", "attorney")
- Topic 5 identifies scams exploiting religious related subjects ("lord", "god", "charity", "life")

3.5 Discussion

The results suggest that the NMF-based topic modeling approach was able to identify coherent and distinct (with some exceptions) themes within the email dataset, despite the highly repetitive nature of fraudulent content. The relatively high diversity score confirms that the discovered topics are well separated, but the coherence score could be brought higher by improving the coherence between each topic's terms. However, the relatively high reconstruction error indicates that the model's approximation of the data could be improved by further tuning the regularization parameters or increasing the number of components. Future work could involve testing alternative vectorization strategies or applying hybrid models combining NMF with clustering approaches to further improve topic separation and model interpretability.

4 Concluding remarks

The experimental results demonstrate that Non-Negative Matrix Factorization (NMF) is an effective technique for extracting meaningful and distinct topics from a dataset of fraudulent emails. The model was able to identify consistent patterns within the dataset, revealing key themes such as financial fraud, inheritance scams, lottery scams, and impersonation of government officials.

The evaluation metrics confirm the quality of the extracted topics. The relatively high coherence score indicates that the terms within each topic are semantically related, suggesting that the discovered topics are internally consistent. However, the reconstruction error remains relatively high, which suggests that the model's ability to capture the underlying data structure could be improved.

The high diversity score indicates that the model successfully identified a wide range of fraudulent patterns, despite the highly repetitive nature of fraudulent emails. For instance, inheritance scams and financial fraud showed high consistency within individual topics, while impersonation and lottery scams were well separated into distinct clusters.

4.1 Limitations

Despite the overall success, there are some limitations to consider:

- The high reconstruction error suggests that the model may not fully capture the variability of the input data. This could be addressed by increasing the number of components or adjusting the regularization parameters.
- The model relies on the bag-of-words representation, which ignores word order and contextual meaning. This limitation could be mitigated by adopting more sophisticated embedding techniques (e.g., Word2Vec, BERT) or hybrid approaches.
- Fraudulent emails often share similar language patterns and terms, which may lead to overlapping topics or misclassification of terms between similar categories.

4.2 Future Work

Future work could explore several potential improvements:

- Applying contextual embeddings (e.g., BERT) to better capture semantic similarities and word relationships.
- Combining NMF with clustering techniques (e.g., K-Means) to further refine the separation of overlapping topics.
- Expanding the dataset to include more diverse samples of fraudulent communication, including other types of scams and phishing attempts.
- Introducing a post-processing step to group similar topics or merge highly correlated topics to improve interpretability.

In conclusion, the use of NMF for topic modeling in the context of fraudulent emails has shown promising results in terms of coherence and topic separation. Fine-tuning the model configuration and exploring hybrid techniques could further enhance the performance and interpretability of the discovered topics.

AI Usage

Parts of this project have been realized with the assistance of the GPT-4o LLM (ChatGPT). In particular:

- collecting ideas about the evaluation of the obtained results
- generation of the regular expressions used in the structural classification (greetings and closings)
- grammar-wise review and correction of this report
- sometimes, as a ”documentation browser” or ”explainer” for Python functions
- help in resolution of various issues in the code. The solutions provided by the AI model were manually reviewed and tested both individually and as part of the full workflow

References

[1] Agrawal, C.K.C.S..G.V. S.: Scalable ad-hoc entity extraction from text collections. Proceedings of the VLDB Endowment **1**(1), 945–957 (2008) https://doi.org/10.14778/1453856.1453958

[2] Al-Moslmi, O.M.G.O.A.L..V.C. T.: Named entity extraction for knowledge graphs: A literature overview. IEEE Access **8**, 32862–32881 (2020) https://doi.org/10.1109/ACCESS.2020.2973928

[3] Hong, C.J.Y.H.K.Y..S.J. T.: Knowledge-grounded dialogue modelling with dialogue-state tracking, domain tracking, and entity extraction. Computer Speech Language **78**, 101460 (2023) https://doi.org/10.1016/j.csl.2022.101460