Text Mining Project

Spam Detection and Topic Modeling

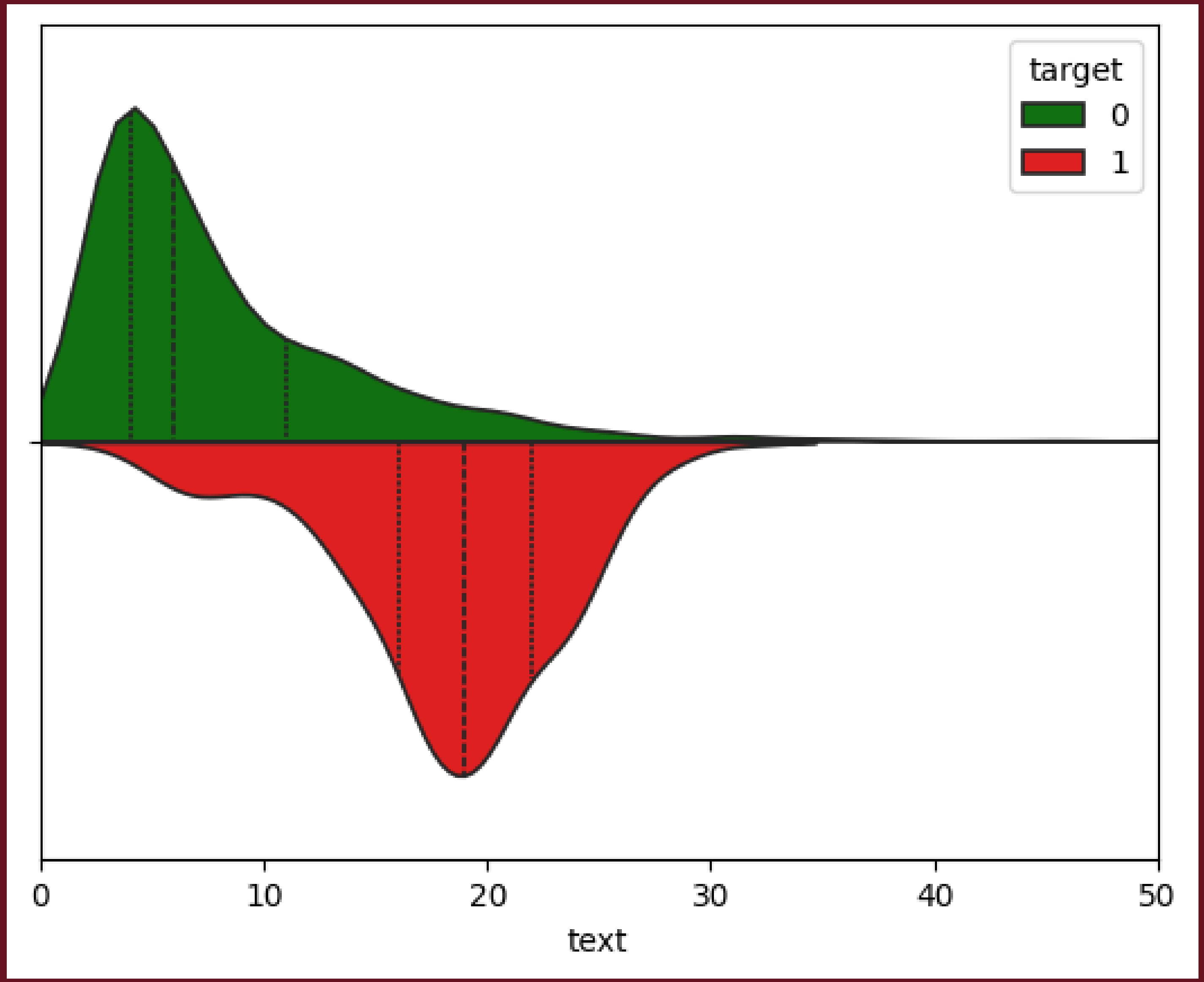
UNIVERSITÀ
MILANO BICOCCA

Castagna Alessandro
Colombaro Daniel
Pasotti Matteo
MSC DATA SCIENCE

Index

Introduction

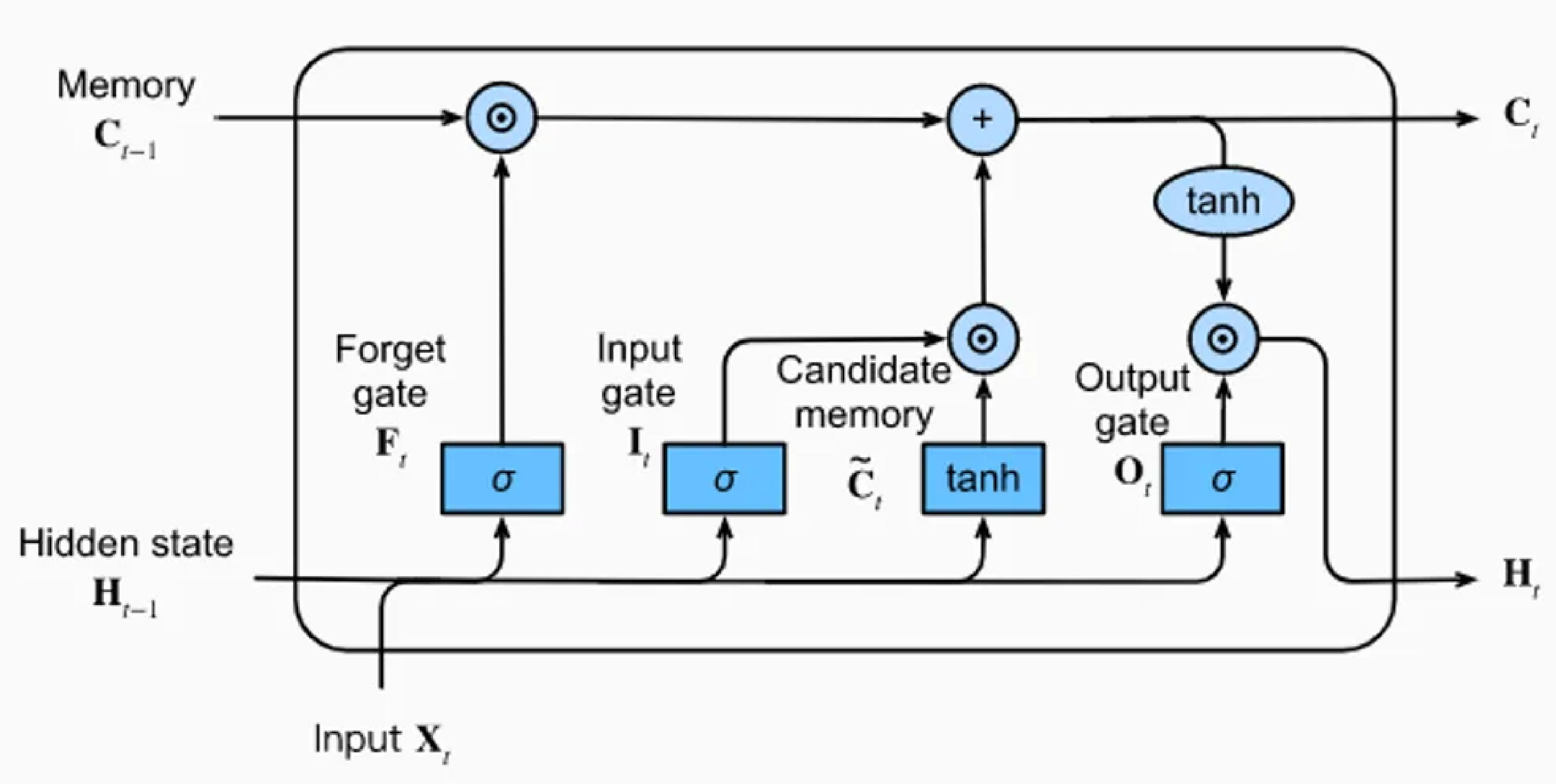
- 5572 messages with two features (“Text” and “Target”)
- 13,4% spam
- TEXT PROCESSING:
  - punctuation removal
  - stopwords removal



Spam Detection Bi-LSTM

A Bi-LSTM consists of two LSTM networks with long dynamic updated by short-term memory and external information coming from previous layer

- **Tokenization:** Keras Tokenizer associating words to vocabulary indices.
- **Embedding:** Keras Embedding layer (NON-contextual).
- **Bidirectional layer:** keeping in memory word embeddings of the rest of the sentence in a dynamic way.



Accuracy: 100% train set - 98.5% test set

<u>**Is this performance reliable?**</u> We compare it with state-of-the-art **BERT transformer**.

Spam Detection BERT

Transfer learning is the practice of using pre-trained models and adapt them to a different task.
BERT uses attention masks to replace 15% of the tokens randomly and let users know which tokens contain real information. It enables contextualized word embedding.

- **Tokenization:** BERT Tokenizer working at sub-word level.
- **Embedding:** Contextualized embedding based on neighborhood capable of **sense disambiguation.**
- **BERT layer:** 12 Encoder layers

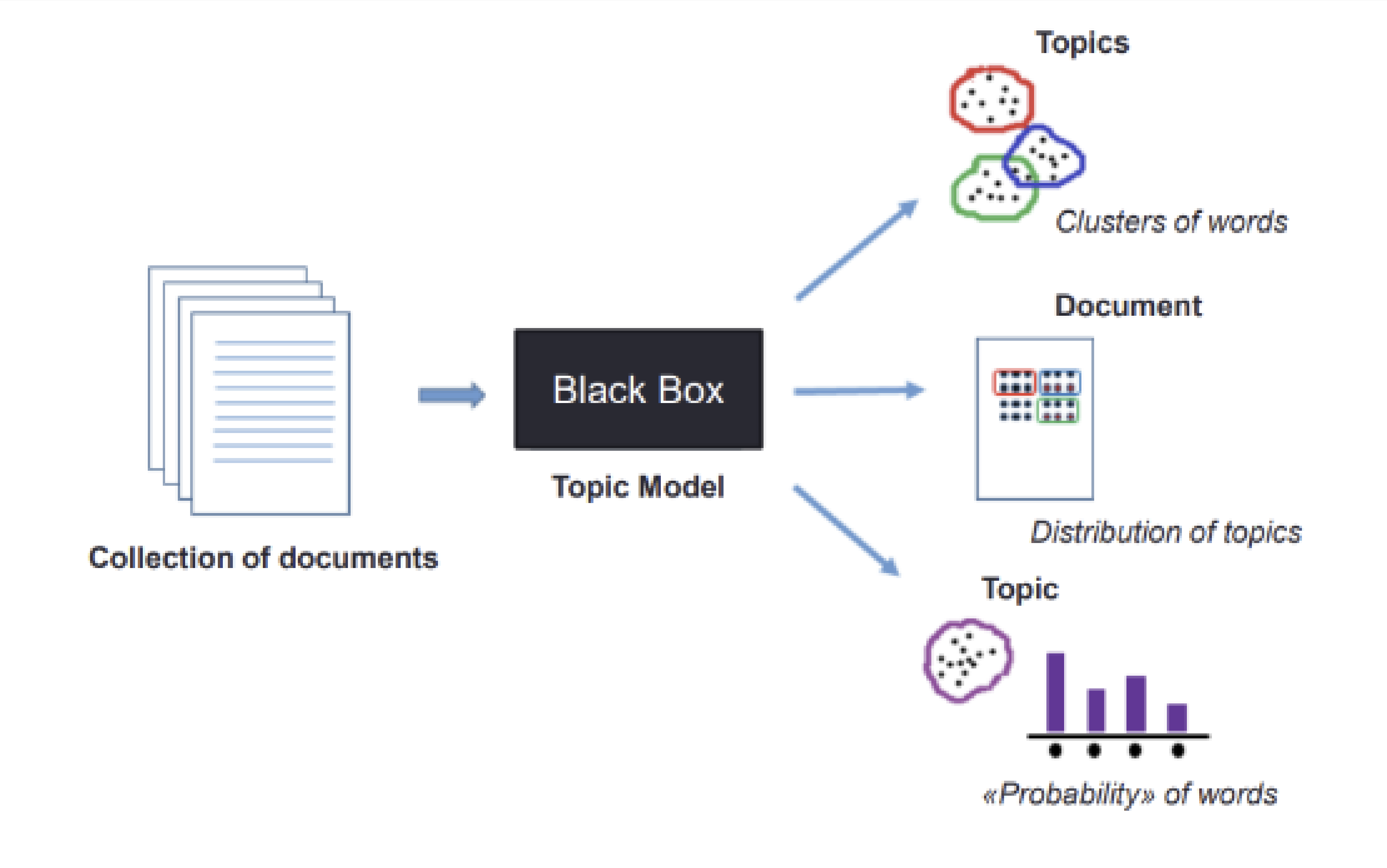
Accuracy: 86.6% train set - 85.6% test set.

Given that the weights of the last layer have not been fine-tuned on the spam detection task, the performance is comparable to Bi-LSTM.

Understanding Topics in Spam Messages Through Topic Modeling

Approach: Applied Latent Dirichlet Allocation (LDA)

Evaluation Metrics:
Coherence: Measures topic interpretability.
Perplexity: Indicates model prediction quality.



Abbreviation Expansion and Choose of the Topic Number

Pattern Observation: Abundant use of abbreviations in email messages, particularly in "ham" messages.

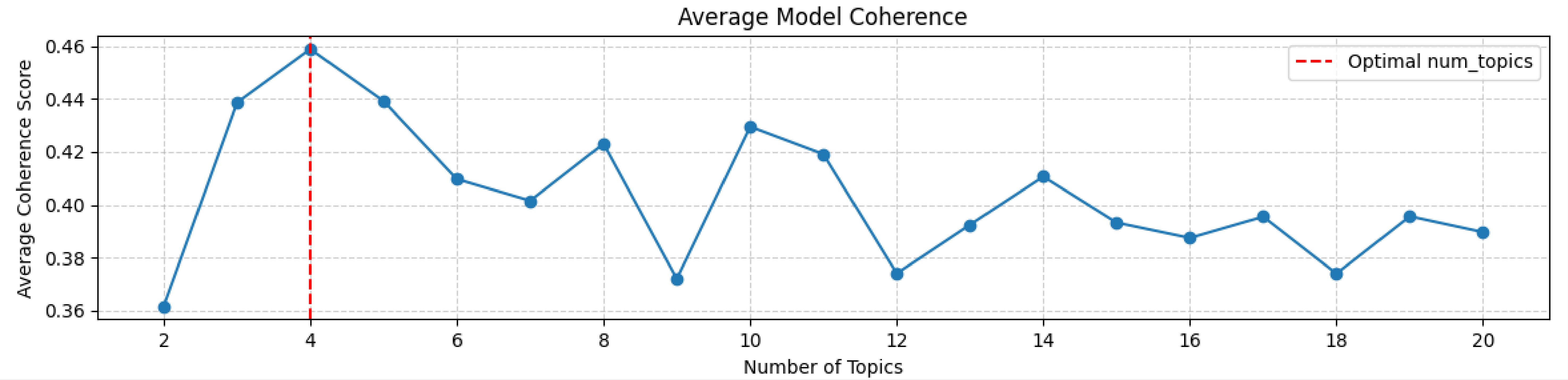
```
Document 3913 (original): yeah whatever lol
Document 3913 (expanded): yeah whatever laugh out loud
```

Approach: Definition of candidate numbers of topics (from 2 to 20 with a step of 2).

Average Coherence values calculated for each number of topics.

Average Model Coherence

Show a line graph with the number of topics on the x-axis and average coherence scores on the y-axis. Highlight a vertical dashed red line at the point where coherence peaks (indicating four topics).



- Model coherence measures how well words within a topic are semantically related.
- Higher coherence indicates that topic keywords form a cohesive and meaningful theme.
- We calculated the average coherence for models with different numbers of topics.
- Topics with high coherence have keywords that relate well to each other

Evaluation

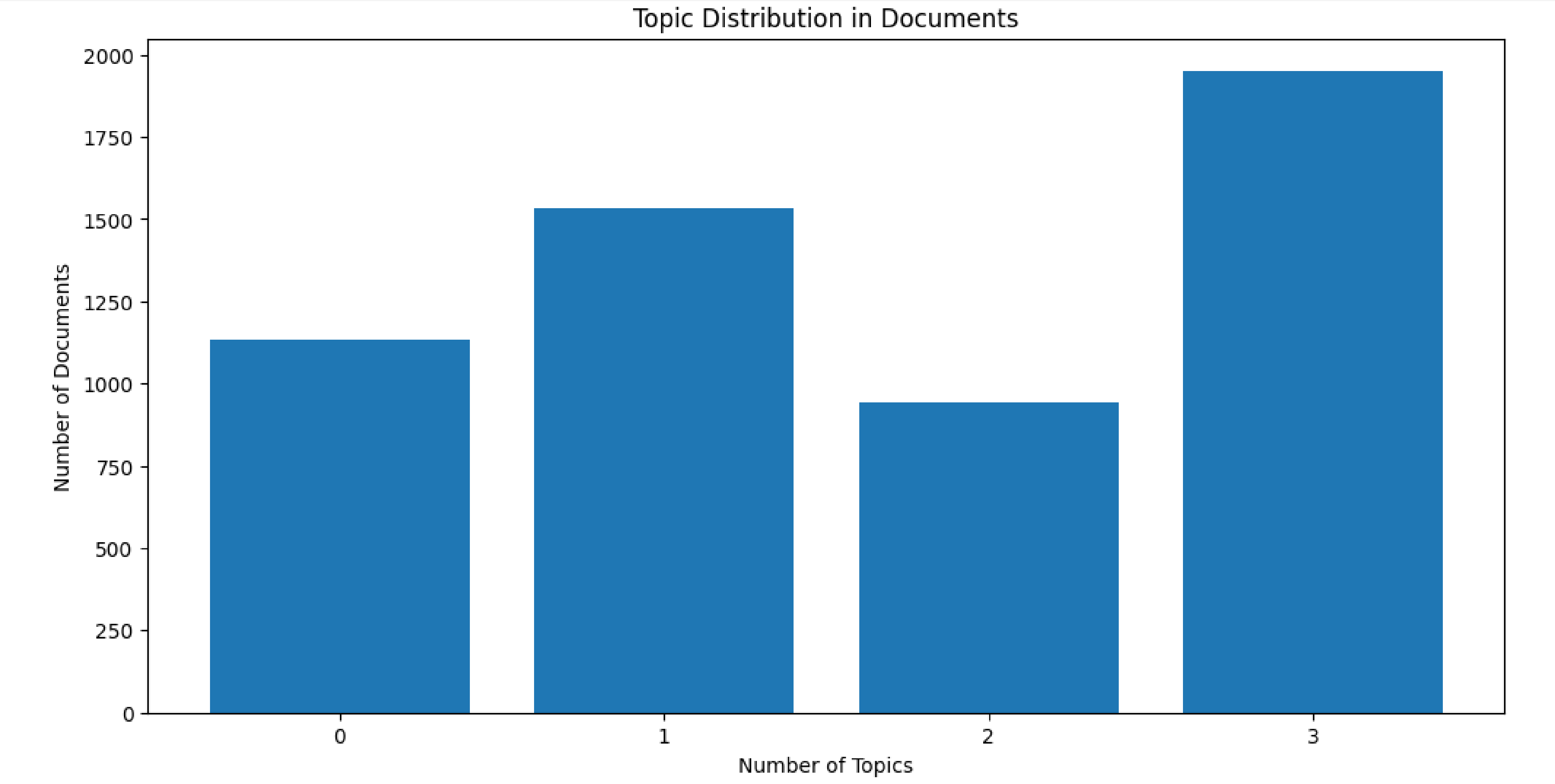
Tools and metrics used: Topic Coherence, Model Perplexity, Topic Visualization (pyLDAvis).

- **Topic Distribution Evaluation:** This involves the analysis of how documents are distributed among identified topics, revealing the prevalence and significance of each theme in the dataset.
- **Topic Coherence:** Topic Coherence is a metric that measures the semantic similarity of words within a topic. It helps assess how well-defined and meaningful the identified topics are.
- **Model Perplexity:** Model Perplexity is a measure of how well a model predicts a sample. Lower perplexity values indicate better model performance in describing the dataset.
- **Topic Visualization (pyLDAvis):** pyLDAvis is a tool used to create interactive visualizations of topic modeling results. It provides insights into the relationships between topics and highlights the most salient terms for each topic.

Evaluation of Topic Distribution

Overview of the Topic Modeling approach and its probabilistic basis

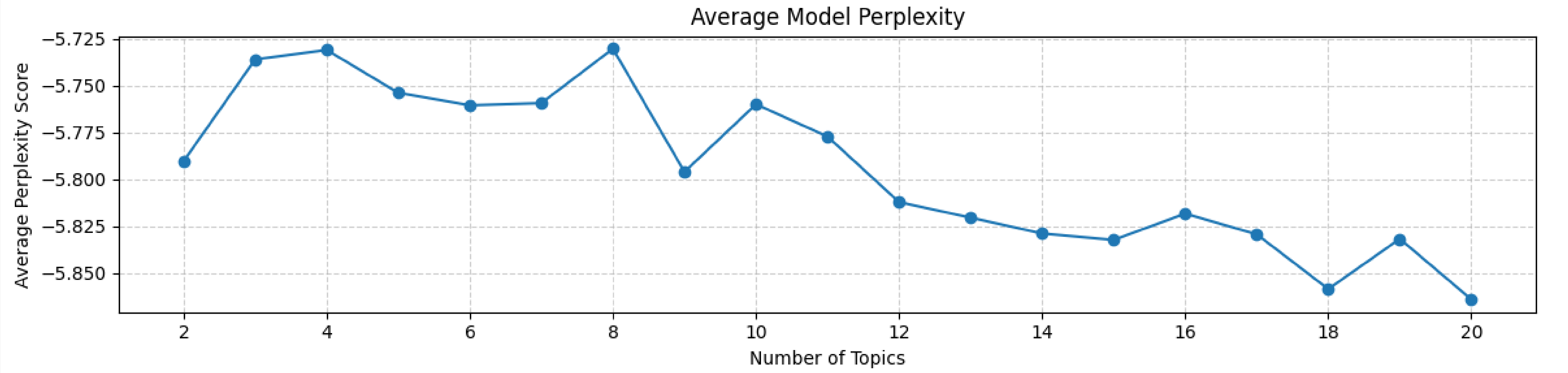
- We used Topic Modeling to uncover themes within our text documents.
- Each document was assigned to the topic with the highest probability, revealing the dominant theme.
- The bar chart illustrates the distribution of documents among topics.
- Topic 3 is highly prevalent, indicating substantial interest or relevance in the dataset.



Display a bar chart with four distinct bars labeled as Topic 0, Topic 1, Topic 2, and Topic 3.

Average Model Perplexity

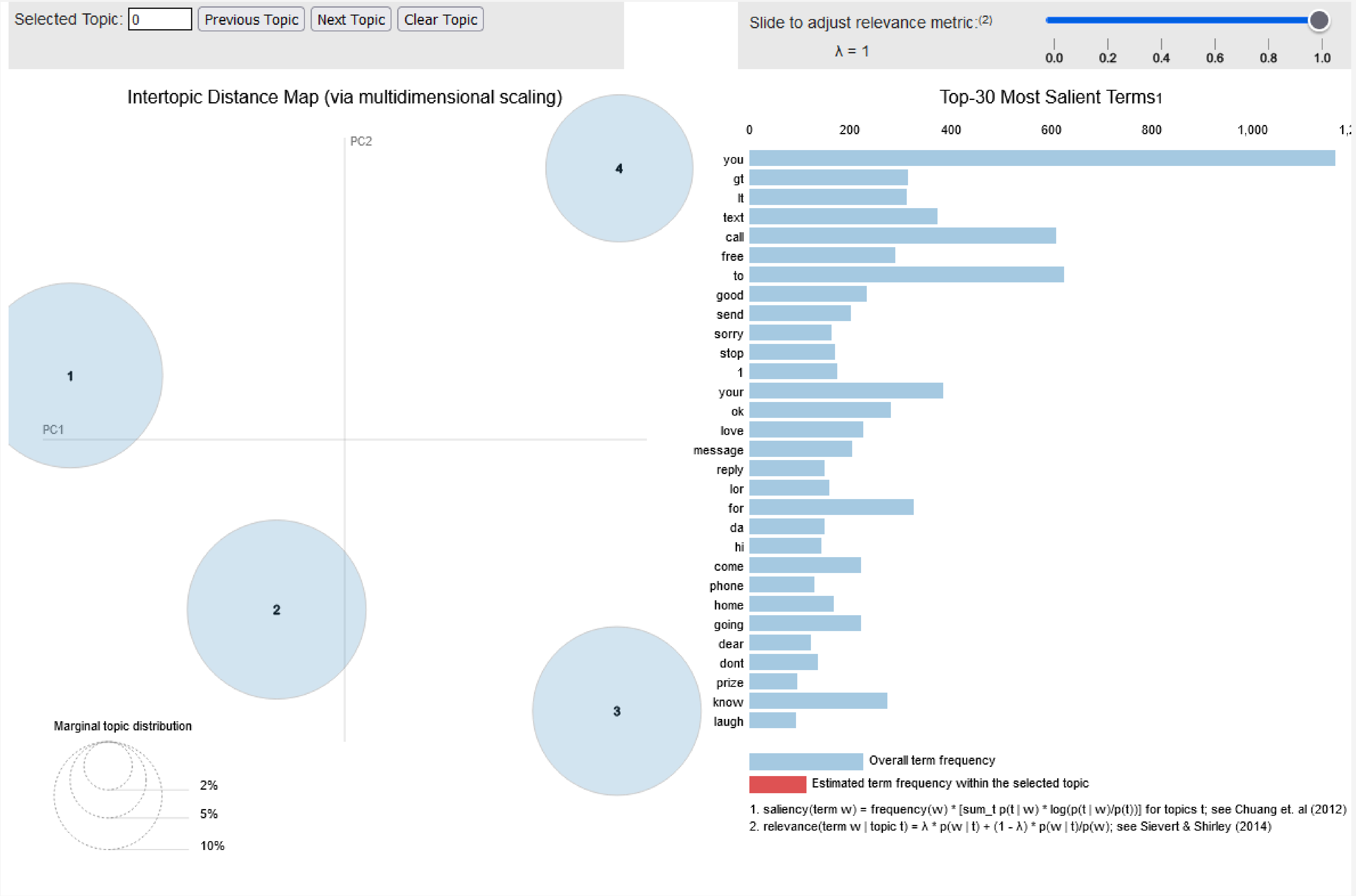
Display a line graph with the number of topics on the x-axis and average model perplexity (negative values) on the y-axis.



- Perplexity measures how well a model predicts a sample, with lower scores indicating better performance.
- While lower (more negative) perplexity is favorable, too many topics may not improve qualitative interpretation (as shown by coherence scores).
- The negative values are a result of the logarithmic transformation and indicate better predictive power.

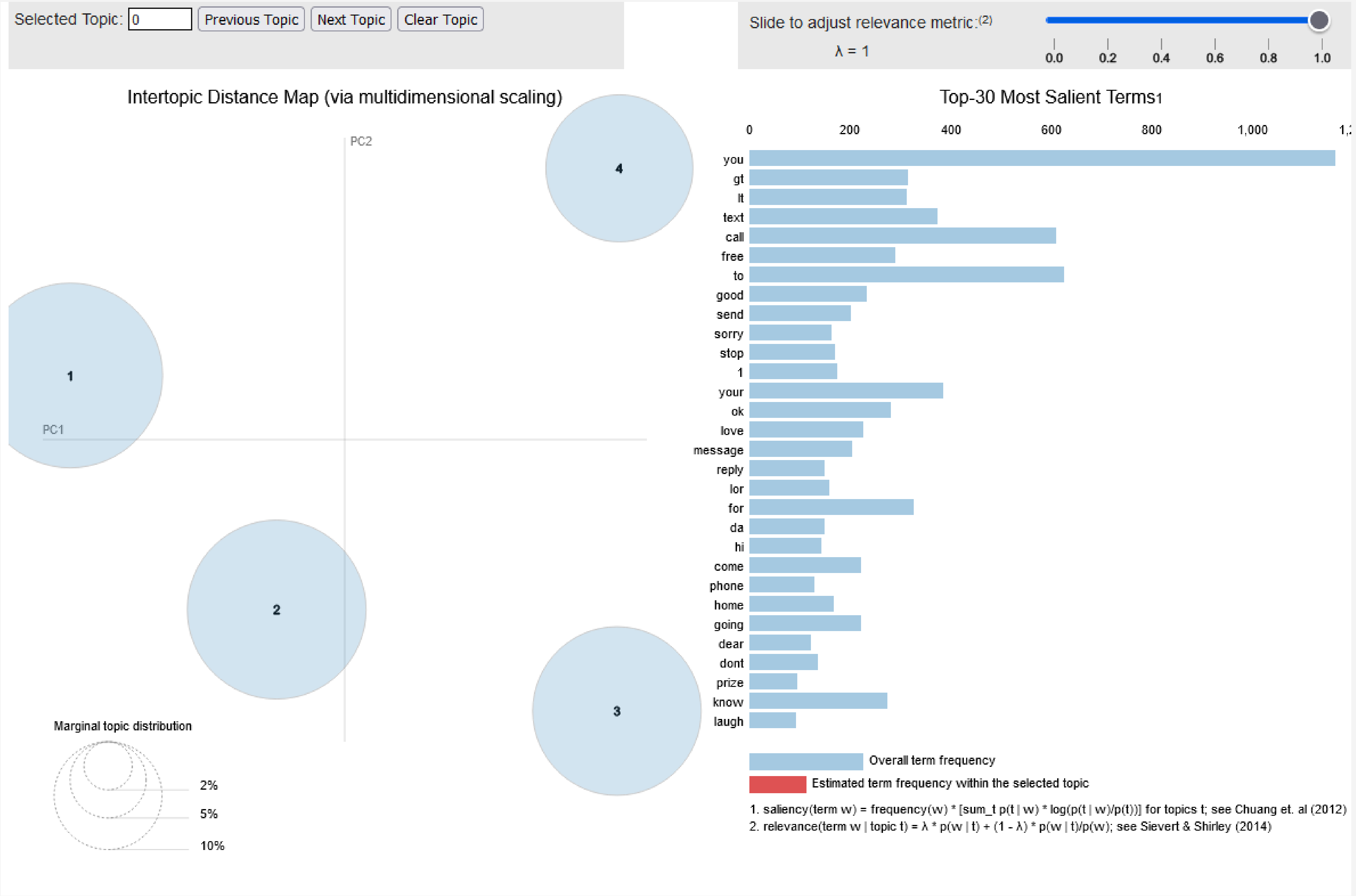
Interactive Topic Visualization with pyLDAvis

- We used the pyLDAvis tool to create an interactive visualization for interpreting Topic Modeling results.
- The visualization consists of two main components:
  - Intertopic Distance Map: Topics represented as circles in a two-dimensional space (PC1 and PC2).
  - Most Salient Terms: A bar chart listing the 30 most relevant terms for a selected topic.



Interactive Topic Visualization with pyLDAvis pt.2

- The Intertopic Distance Map visually showcases relationships between topics.
- Most Salient Terms reveal key words for each topic.
- Relevance metric balances term significance within a topic with their frequency across the entire corpus.
- Allows effortless exploration of data and understanding of topic distinctions.
- Spatial separation on the map indicates thematic differences among topics.



Recap and conclusions

Spam Detection

Manageable task by the implemented model.

Performance difference due to fine-tuning of weights.

Further Developments:

Identifying relevant tokens for the specific task

Fine-tuning the last BERT layer

Transfer learning on state-of-the-art LLMs: LLama2.0, GPT-4, Google Bard

Topic Modeling

Effective LDA model for theme identification.

Low perplexity and high coherence scores ensure model accuracy.

PyLDAvis tool enriches topic visualization and comprehension.

Further Developments:

Refine abbreviation expansion list to align with linguistic trends.

Include temporal analysis for dynamic topic evolution tracking.