Laura Claudia Verdesca
24/01/2025

Further analysis on the thesis of Megi Ceka titled From text to insights: **Leveraging Topic Modeling to Explore Climate Change’s Impact on Cultural Heritage Literature**

Index

- Introduction
- Probability distribution
- Classification
    - Supervised Classification
    - Neural networks
    - BERT
    - Ensemble method
- Model testing with new dataset
- Overtourism
    - Topic models application
- Conclusions

Climate change's impact on cultural heritage

- Cultural heritage is the **legacy of physical artifacts,** such as monuments, art pieces, archaeological sites, and **intangible practices,** such as traditions and rituals, that are inherited from past generations, maintained in the present, and passed down to future generations.
- Although climate change is a widely debated topic, its connections to and **harmful effects on cultural heritage** are still not explored.
- This knowledge gap must be carefully bridged by analyzing a **vast body of literature**, in order to identify predominant themes and topics that encompass the entire, multifaceted problem.

Topic Models

Topic Modeling is a subfield of **Natural Language Processing (NLP)**, and enables the analysis of large datasets, uncovering underlying topics and themes within the text.

A topic in topic modeling is essentially a **cluster of words** that frequently occur together. For instance, in a collection of news articles, a topic might consist of words like "economy", "stocks", "market", and "trade" representing a theme related to finance.

The following techniques were applied to a dataset of **850 papers:**

- **LDA** (Latent Dirichlet Allocation)
- **LSI** (Latent Semantic Indexing)
- **NMF** (Non-Negative Matrix Factorization)
- **HDP** (Hierarchical Dirichlet Process)
- **STM** (Structural Topic Model)
- **CTM** (Correlated Topic Model)

Megi's results

Technique	N° of Topics
LDA	5
LSI	4
HDP	8
NMF	3
STM	6
CTM	6

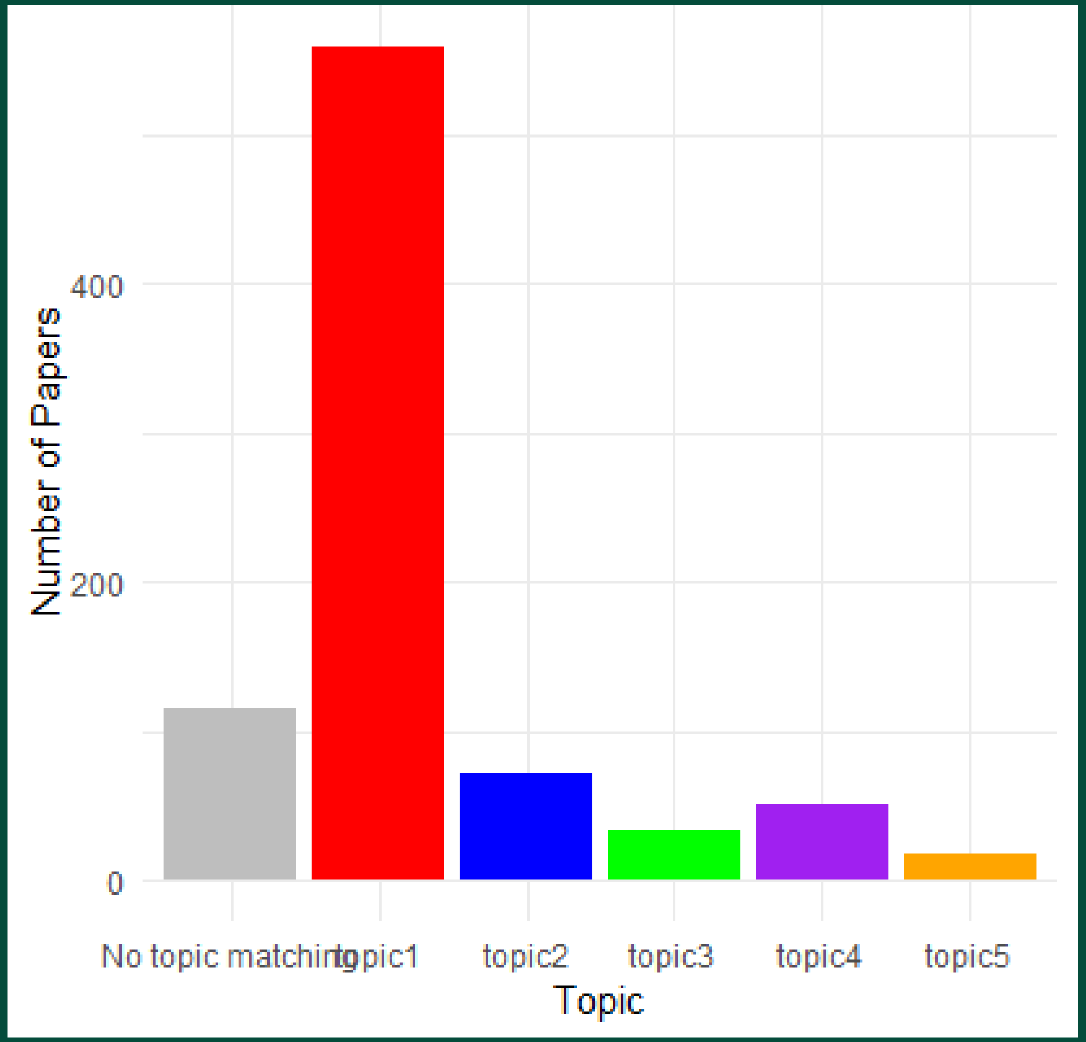
Most common topics	Techniques
Climate Risks to Heritage Sites	LDA, LSI, HDP, STM
Indoor Environmental Quality in Museums	LDA, LSI, NMF, STM, CTM
Climate Change and Archaeological Sites	LDA, HDP, CTM
Indigenous Adaptation Strategies	LDA, HDP, STM
Flood Hazard and Vulnerability Assessment	LSI, NMF, HDP
Urban Flood Risk	LSI, CTM, HDP
Air Pollution and Material Conservation	NMF, STM, CTM
Risk and Hazard Assessment	STM, CTM, HDP

Probability retrieval and distribution

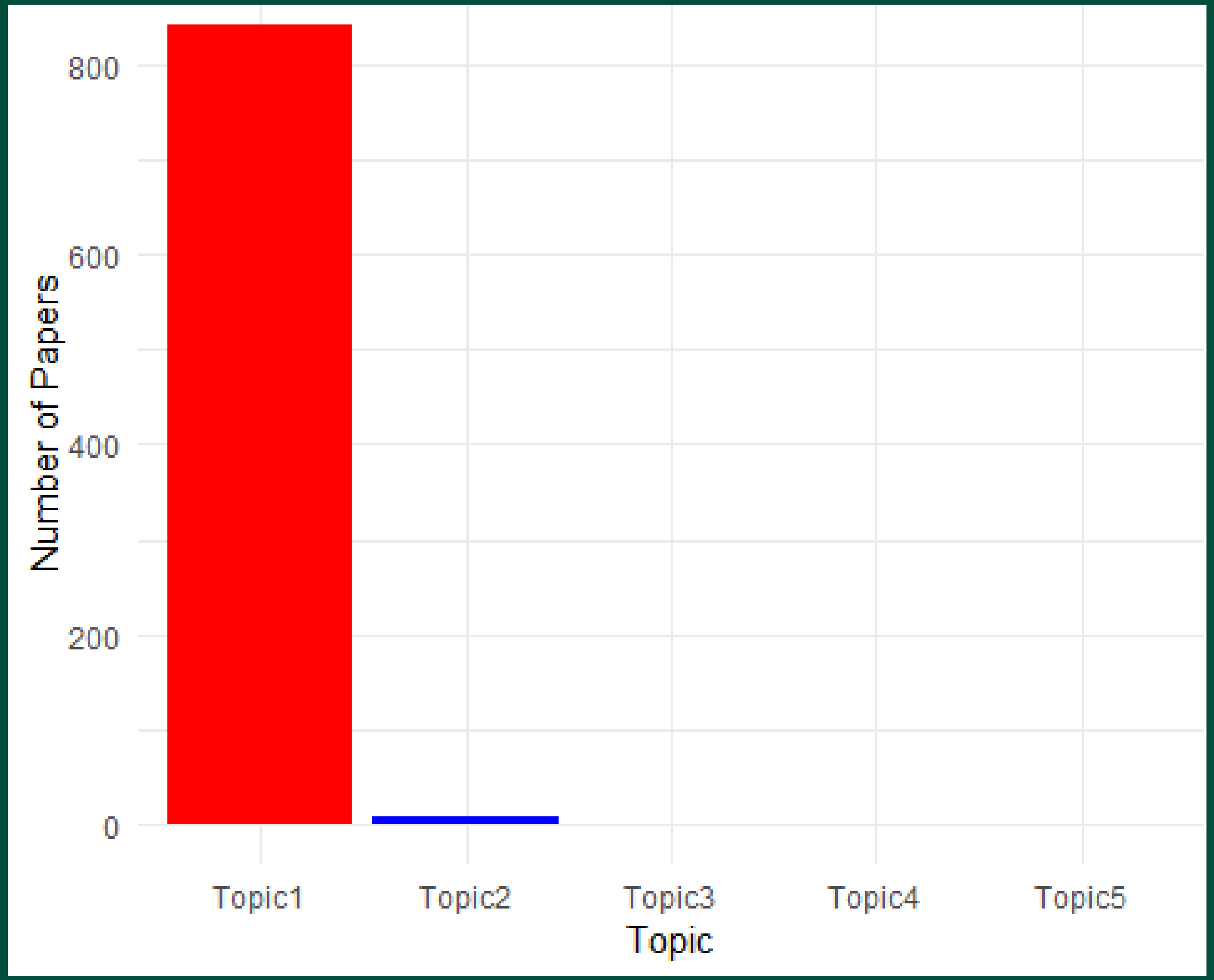
We implement two approaches to retrieve the weight of each topic in each document in the dataset and map the most dominant topic to each paper:

- **Heuristic approach:** each topic is assigned based on how many words match the corresponding word cloud to the abstract and title, the dominant topic is then identified
- **Exact approach:** for each document a probability or "weight" distribution is computed, the dominant topic is then identified

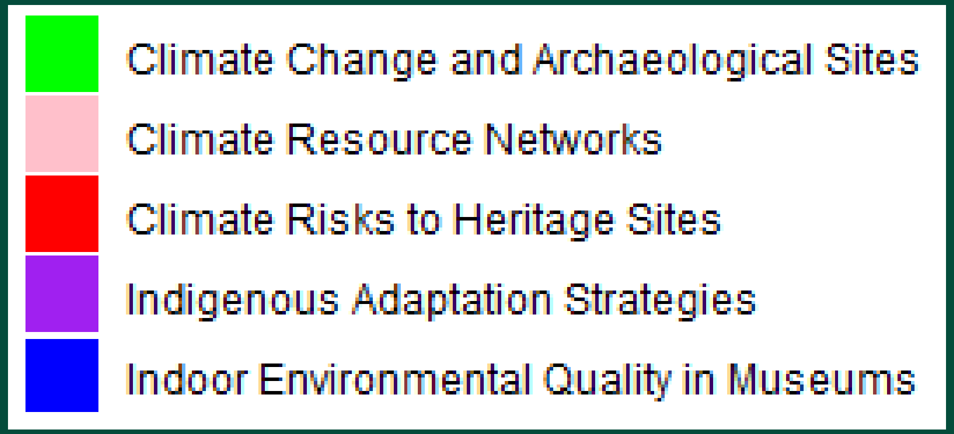
LDA



Heuristic approach

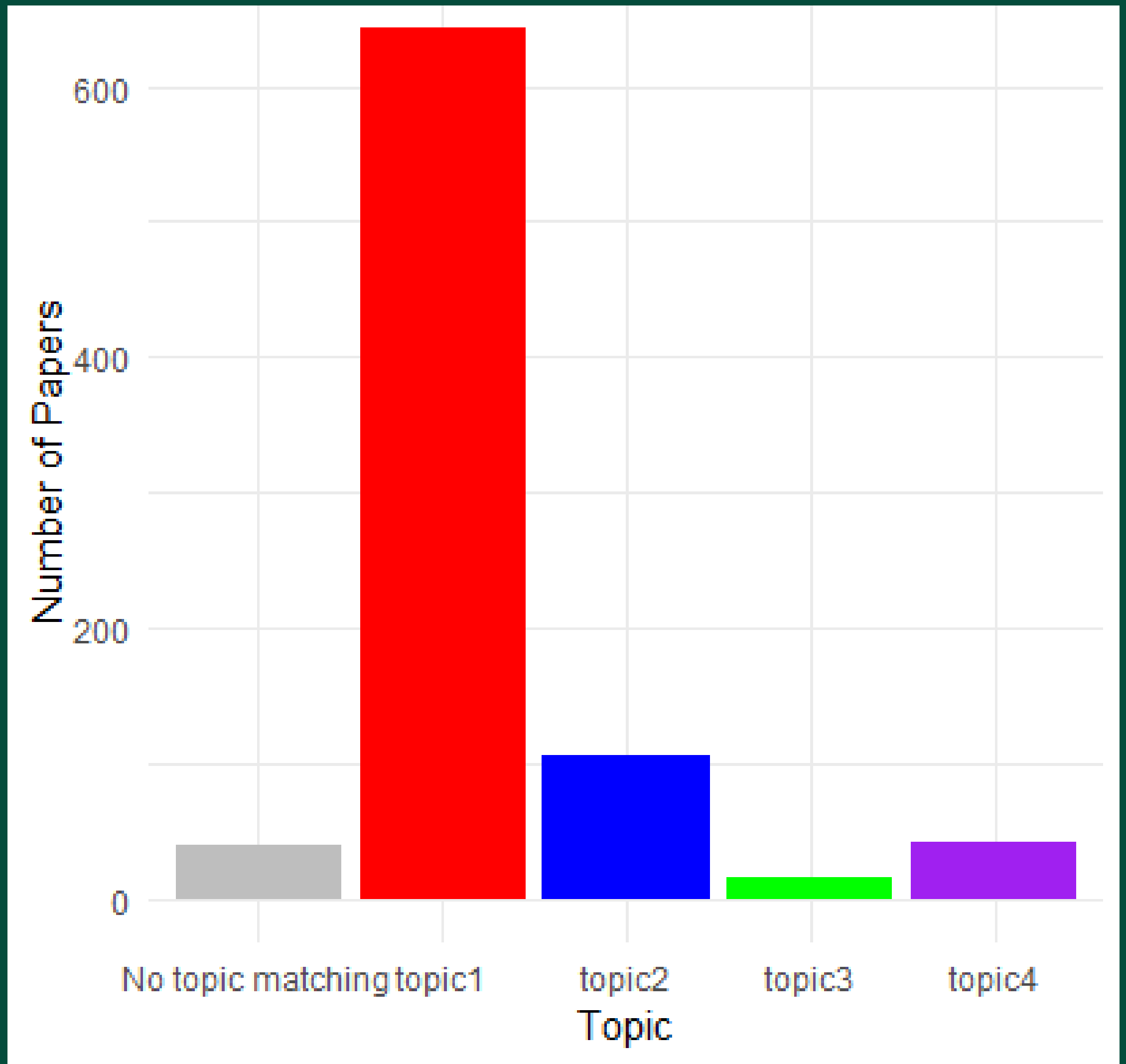


Exact approach

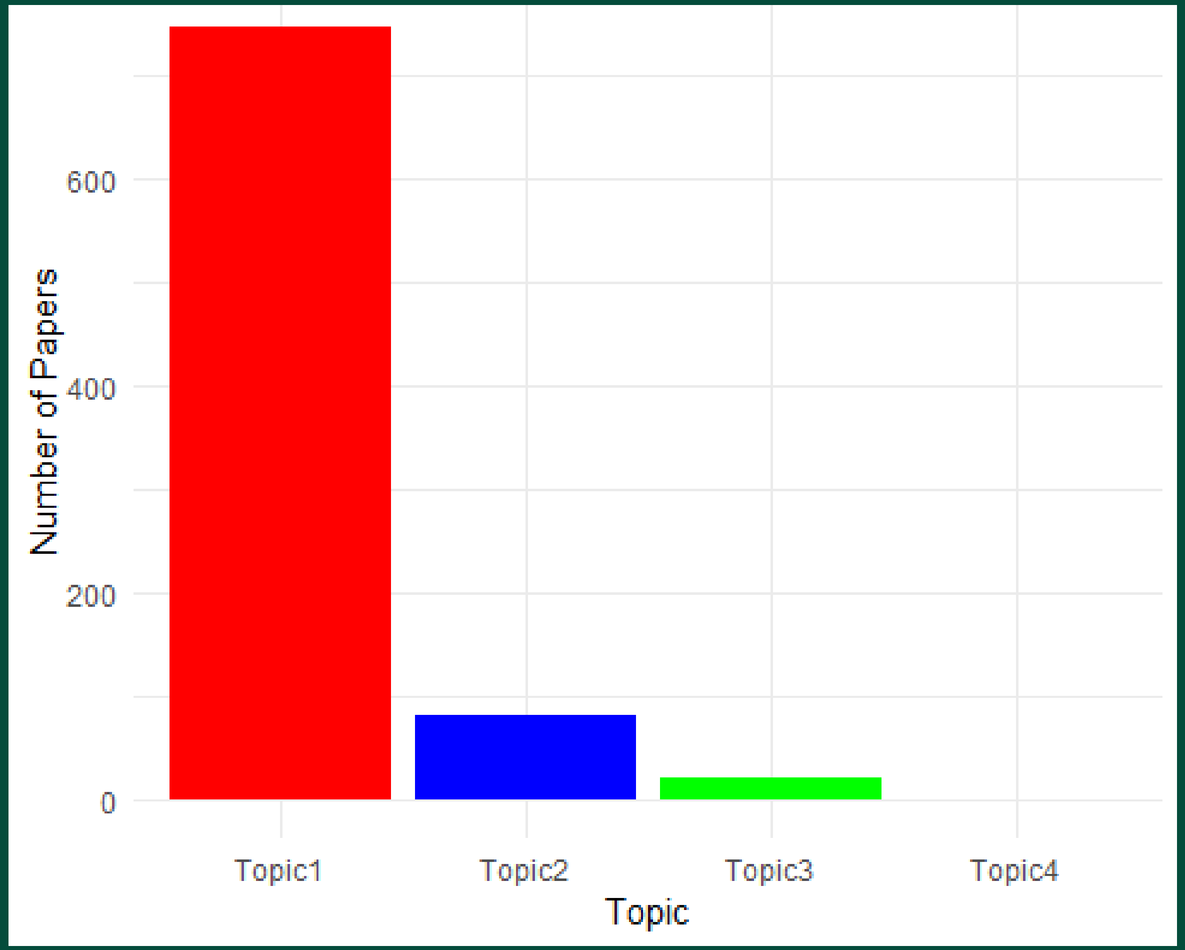


The distribution is quite similar, with Topic 1 being the most dominant one in both cases. Topic 1 corresponds to "Climate Risks to Heritage Sites".

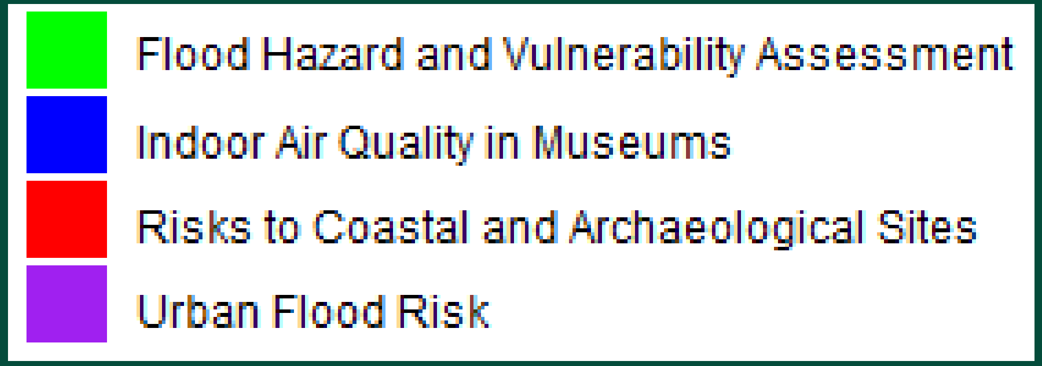
LSI



Heuristic approach

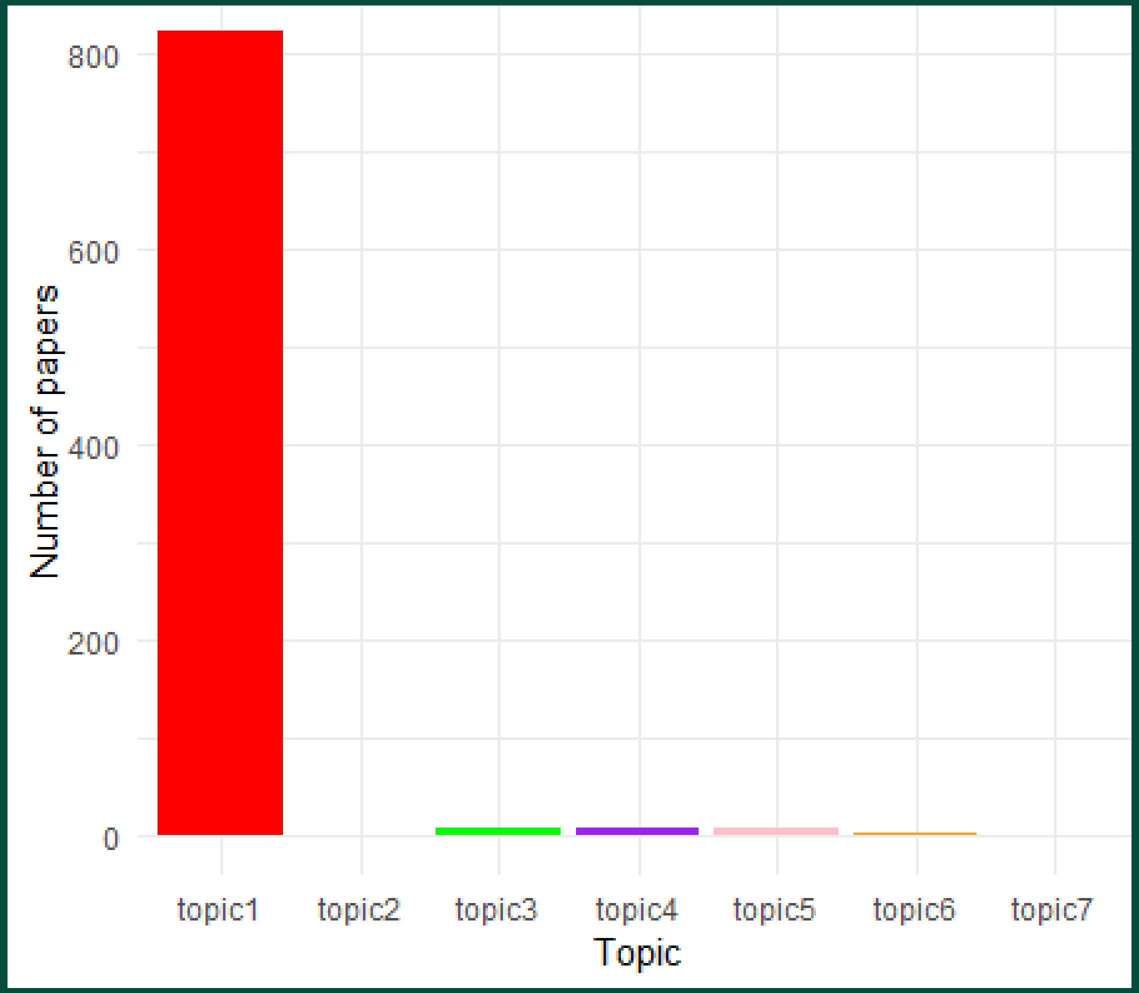


Exact approach

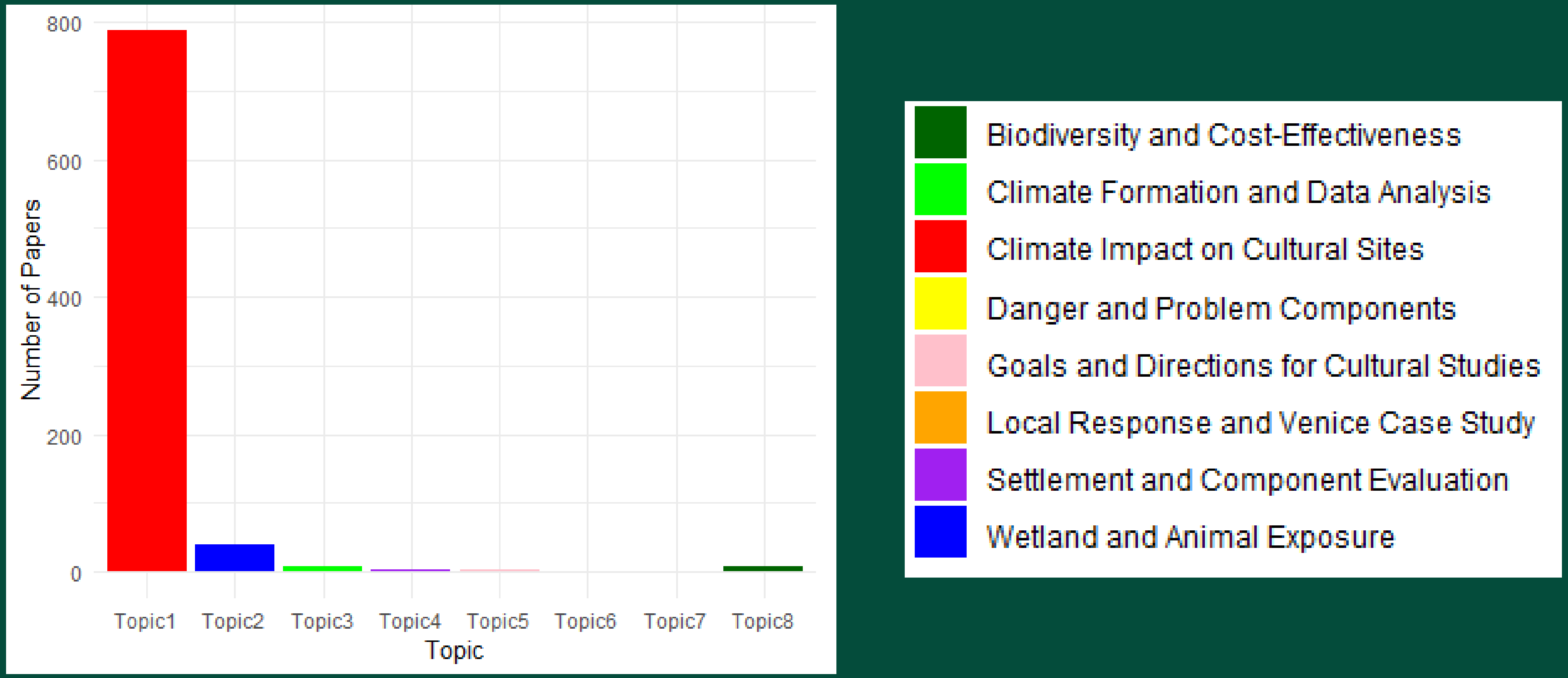


The distribution is quite similar, with Topic 1 being the most dominant one in both cases. Topic 1 corresponds to "Risks to Coastal and Archaeological Sites".

HDP



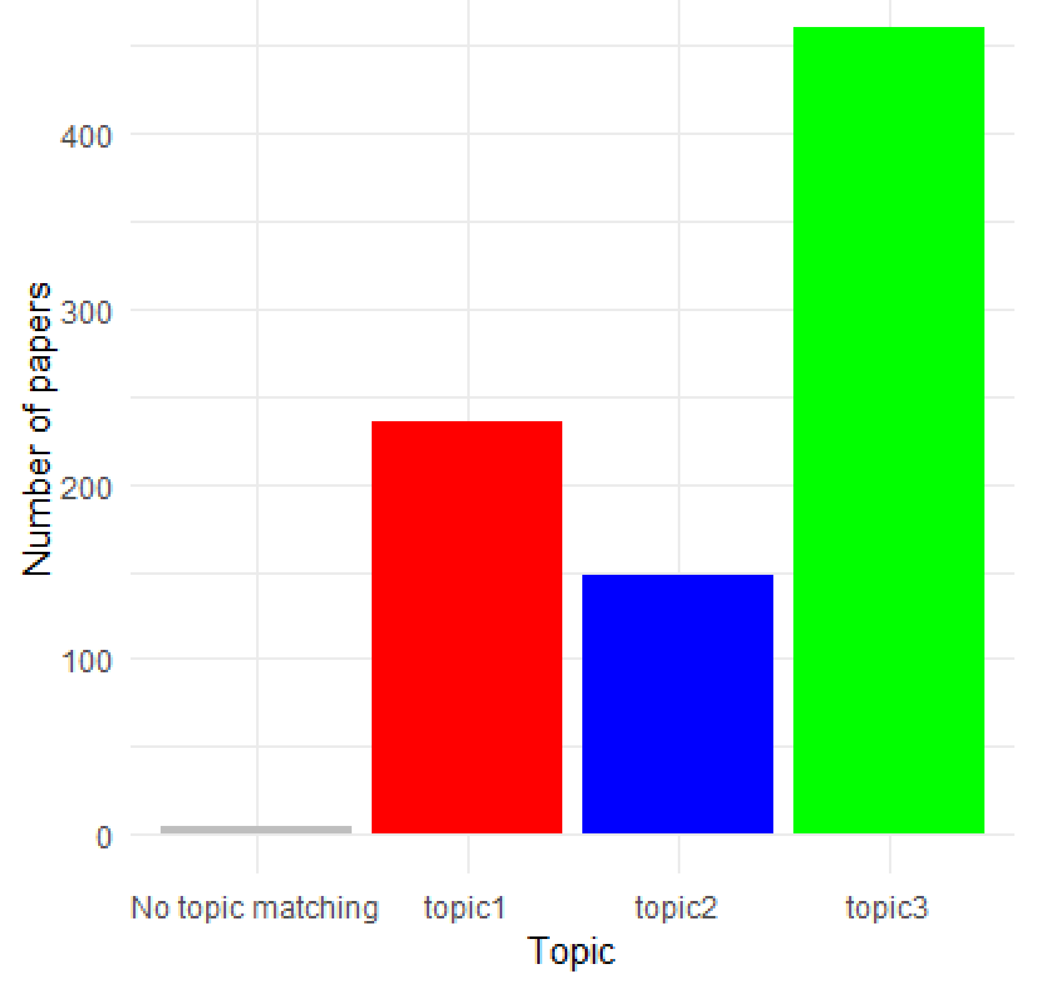
Heuristic approach



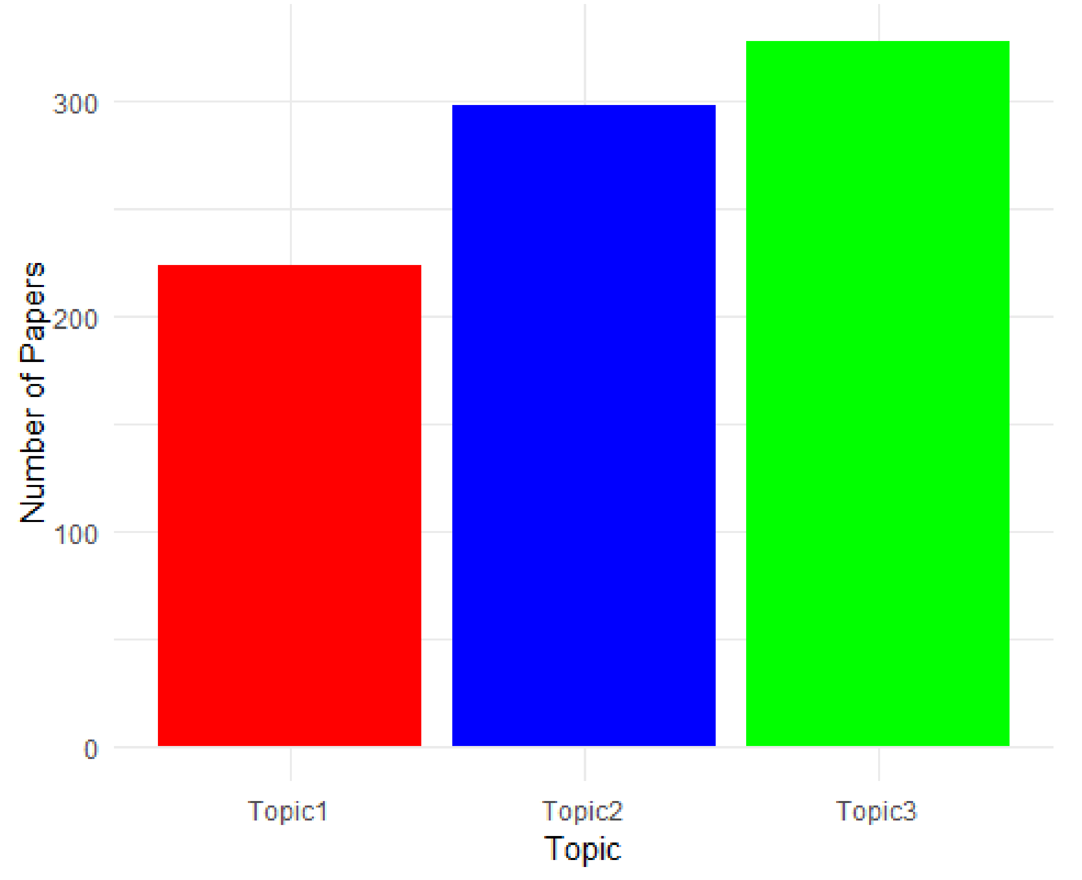
Exact approach

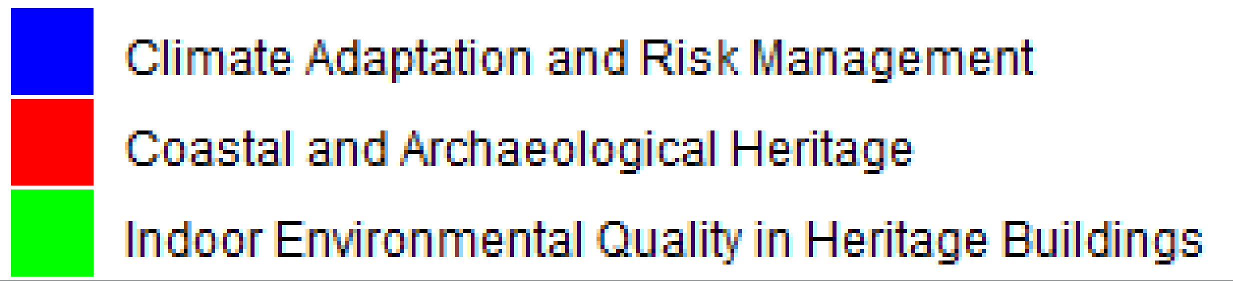
The distribution is quite similar, with Topic 1 being the most dominant one in both cases. Topic 1 corresponds to "Climate Impact on Cultural Sites".

NMF



Heuristic approach

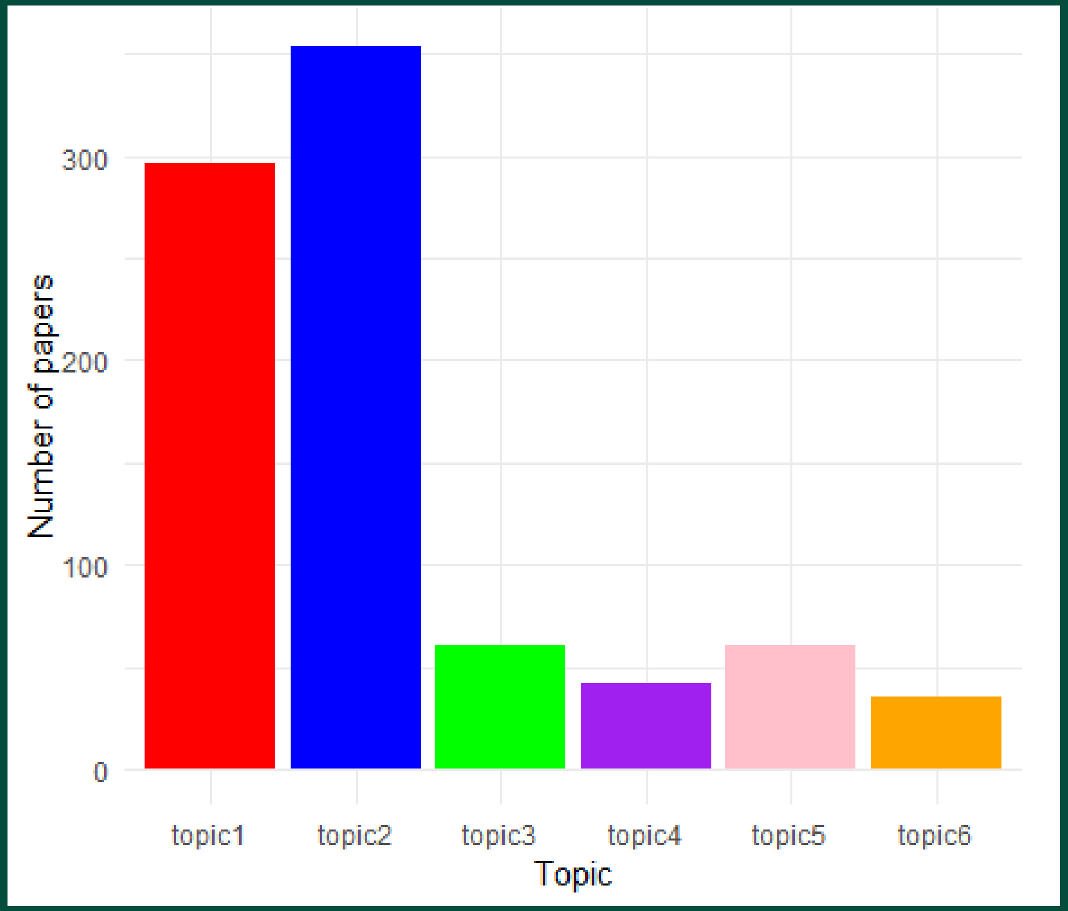


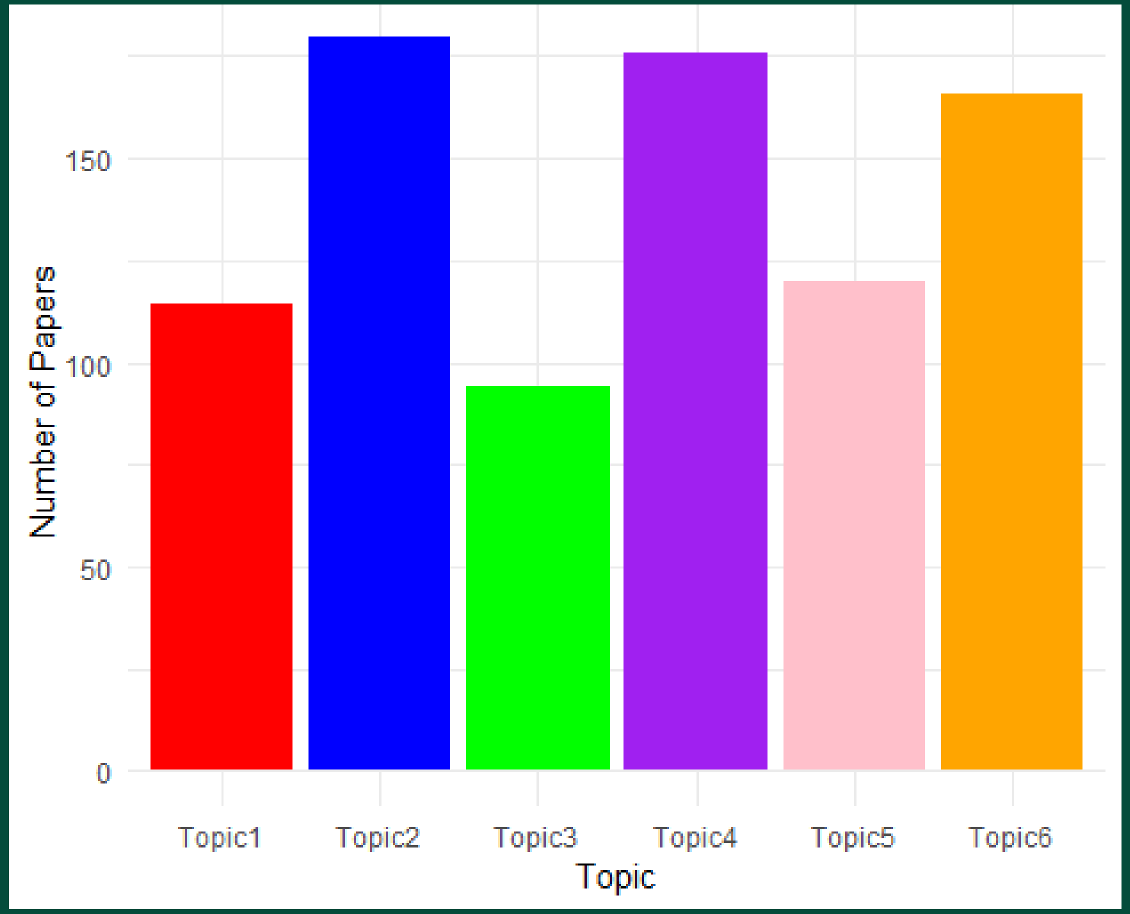


Exact approach

In both cases, the distribution of dominant topics is pretty homogeneous across classes. Topic 3 corresponds to "Climate Adaptation and Risk Management".

STM



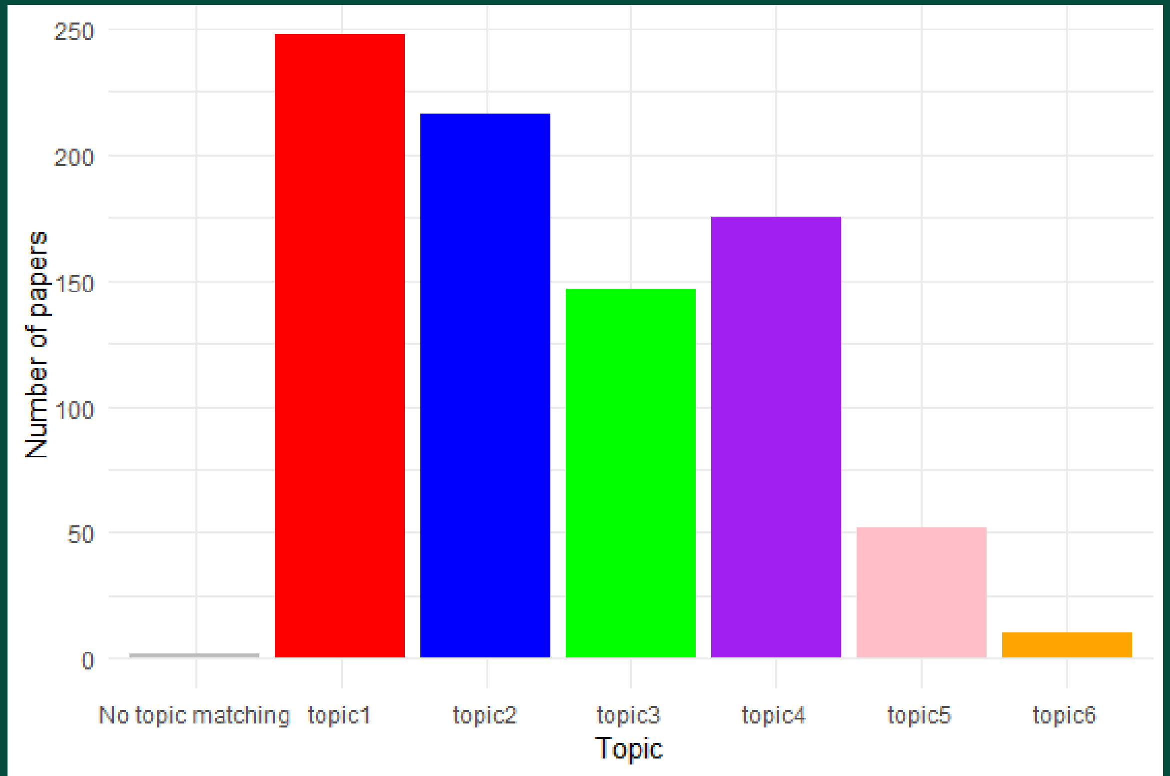


Heuristic approach

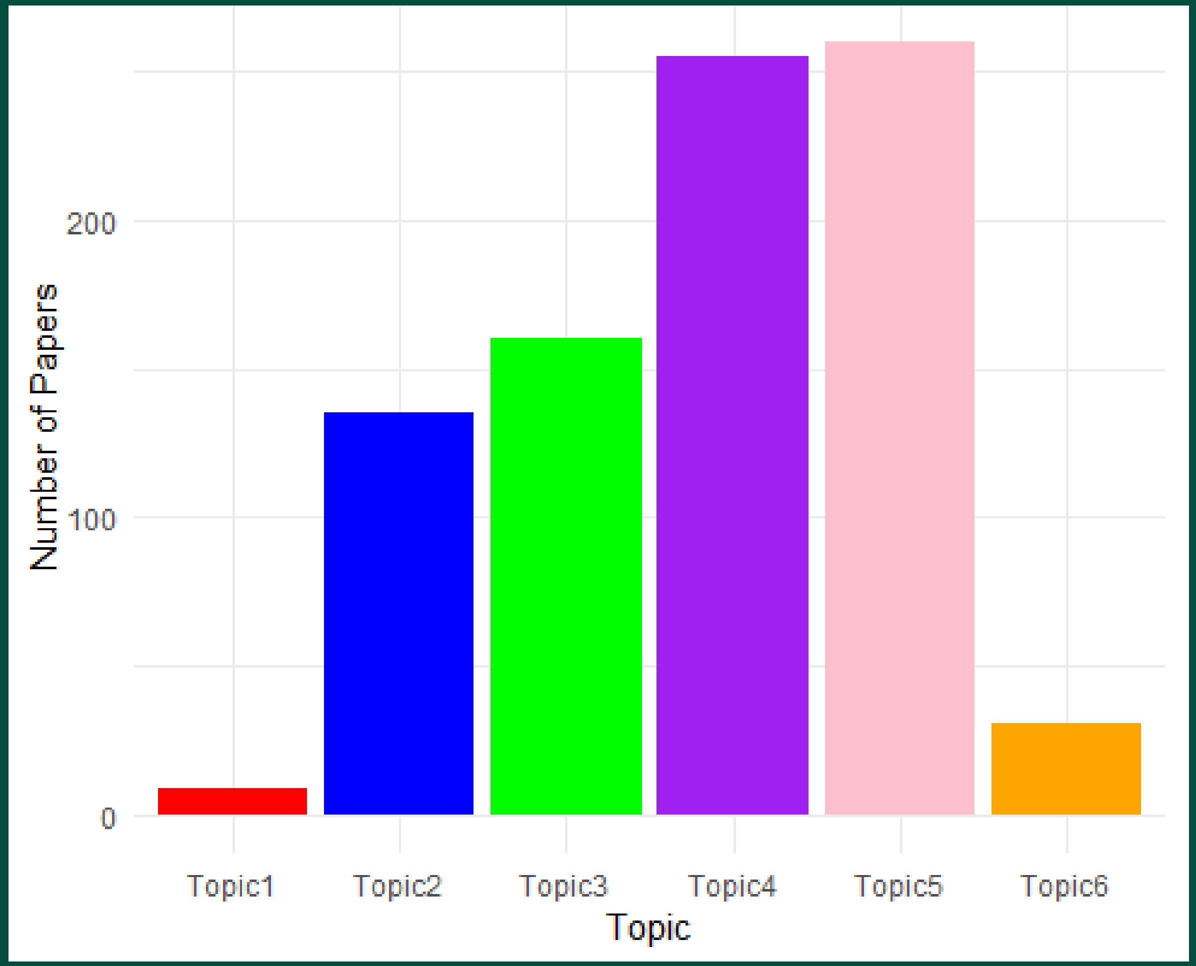
Exact approach

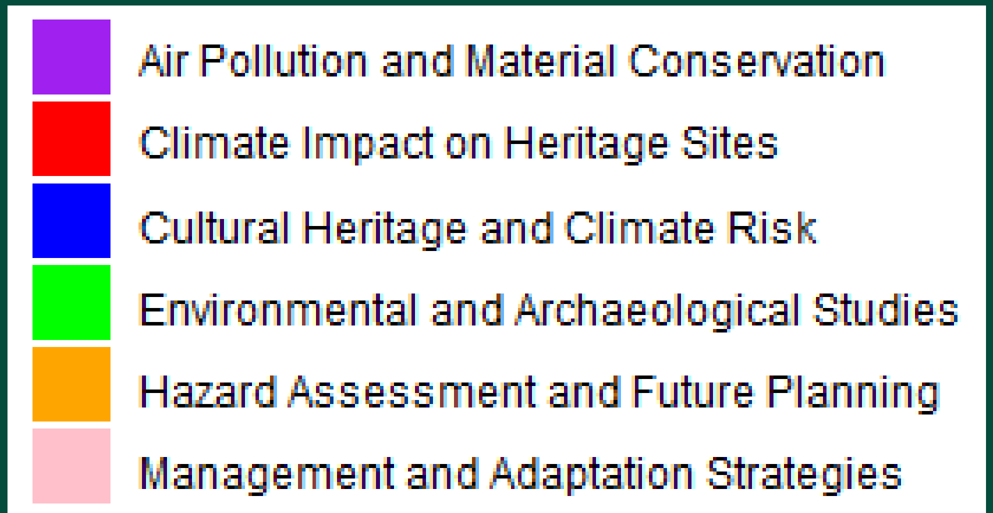
The distribution of dominant topics in the brute approach is underestimated for Topics 3-6, as they show much higher numbers in the second approach. Topics 2-4-6 correspond to “Climate Change Adaptation for Cultural Heritage”, “Air Pollution and Material Conservation”, and “Risk and Hazard Assessment for Cultural Sites”.

CTM



Heuristic approach





Exact approach

The distributions produced by the two approaches look quite different, only Topic 3 and Topic 6 are represented in the same way. Topic 4 and 5 correspond to "Air Pollution and Material Conservation" and "Management and Adaptation Strategies".

Classification

Later we explored the task of classification in order to build a classifier that will allows us to:

- categorize our dataset,
- reveal new patterns and insights
- automatically classify new papers into the appropriate topics without manual intervention

The models we decided to employ are:

- Supervised classification
- Unsupervised classification
- Multi-label classification
- Neural Networks
- BERT
- Ensemble Methods

Supervised Classification

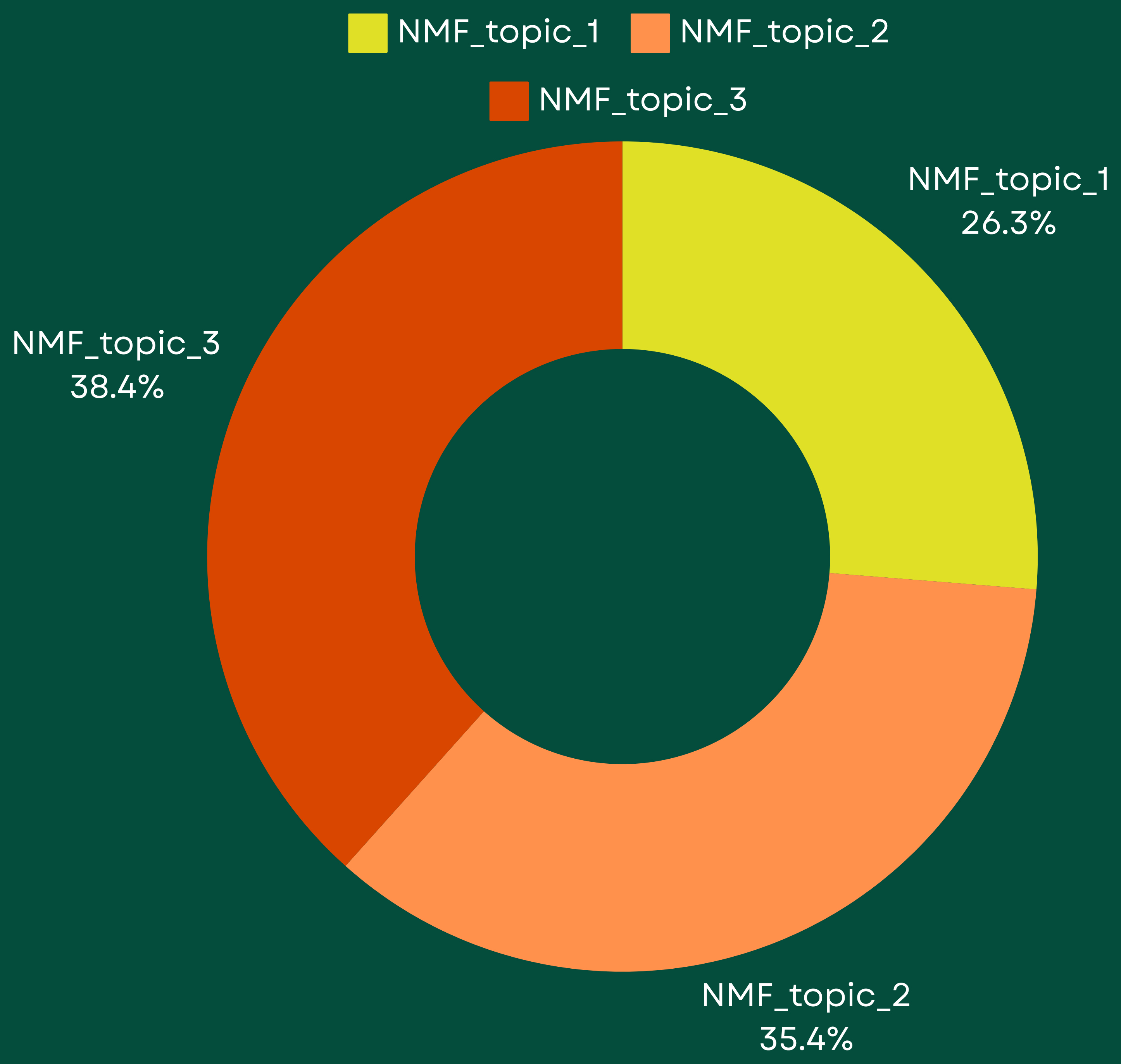
For the supervised models we constructed a **feature matrix** as vectors of length 32, where each entry represents the probability of a topic from each model.

For the **class labels**, we applied various methods to select from these six dominant topics.

We used three different i**mportance functions:**

- Importance function based on normalized frequency
- Manually curated topics
- Importance based on cosine similarity

Supervised Classification: Normalized frequency



The resulting classificaton report:

Topic	Precision	Recall	F1-Score	Support
NMF_topic_1	0.95	0.98	0.96	42
NMF_topic_2	0.98	0.97	0.97	60
NMF_topic_3	1.00	1.00	1.00	68
Accuracy		0.98		170
Macro Avg	0.98	0.98	0.98	170
Weighted Avg	0.98	0.98	0.98	170

Supervised Classification: Manually Curated topics

Class labels	Description
Topic 1	Climate Risks to Heritage Sites
Topic 2	Indoor Environmental Quality in Museums
Topic 3	Indigenous Adaptation Strategies
Topic 4	Flood Hazard and Vulnerability Assessment
Topic 5	Risk and Hazard Assessment

The resulting classificaton report:

Topic	Precision	Recall	F1-Score	Support
Topic 2	0.98	0.98	0.98	51
Topic 3	0.92	0.96	0.94	23
Topic 4	0.93	0.76	0.84	17
Topic 5	0.96	0.99	0.97	79
Accuracy		0.96		170
Macro Avg	0.95	0.92	0.93	170
Weighted Avg	0.96	0.96	0.96	170

Supervised Classification: Cosine Similarity

Cosine similarity measures the similarity between two non-zero vectors by calculating the cosine of the angle between them. In the context of topic modeling, these vectors represent word frequency distributions.

We used them to identify most similar topics across the dataset and unify them, we obtained:

- ’Climate Risks to Heritage Site’
- ‘Indoor Environmental Quality in Museums’
- ’Flood Hazard and Vulnerability Assessment’

The resulting classificaton report:

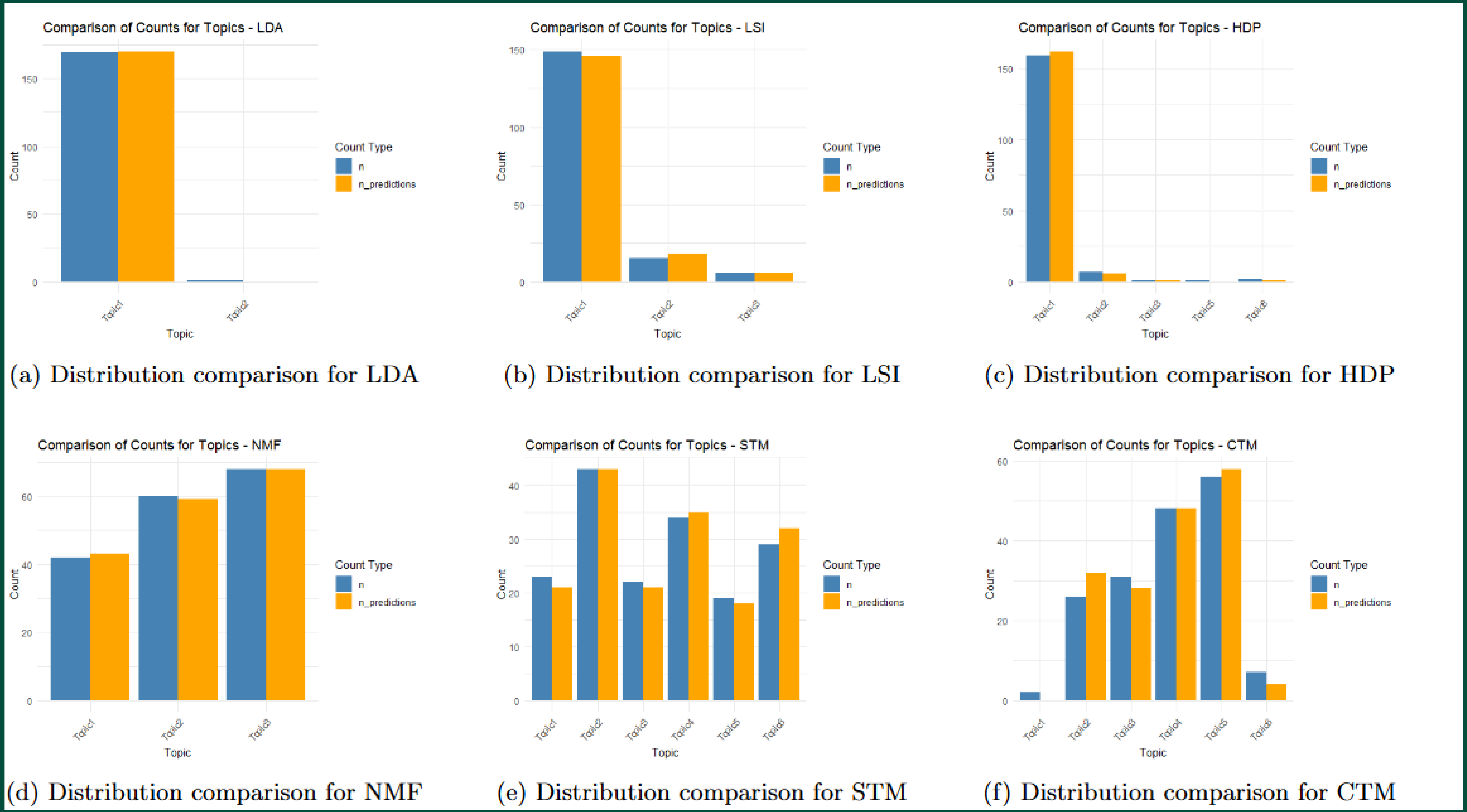
Topic	Precision	Recall	F1-Score	Support
Climate Risks to Heritage Sites	0.98	0.98	0.98	129
Flood Hazard and Vulnerability Assessment	1.00	0.80	0.89	5
Indoor Environmental Quality in Museums	0.95	0.97	0.96	36
Accuracy		0.98		170
Macro Avg	0.98	0.92	0.94	170
Weighted Avg	0.98	0.98	0.98	170

Multi-label Classification

We used as class labels a vector of length 6 to represent each dominant topic for each topic modelling technique.

The resulting classificaton report:

Metric	Value
Hamming Loss	0.0202
F1 Score (Micro)	0.9786
F1 Score (Macro)	0.9323



(a) Distribution comparison for LDA (b) Distribution comparison for LSI (c) Distribution comparison for HDP

(d) Distribution comparison for NMF (e) Distribution comparison for STM (f) Distribution comparison for CTM

Neural Networks

We implemented **Neural Networks models** to see if we could improve the performance of our previous classifiers.

For each classification model we employed a **feedforward neural network (FNN),** also known as a multilayer perceptron (MLP), designed for multi-class classification.

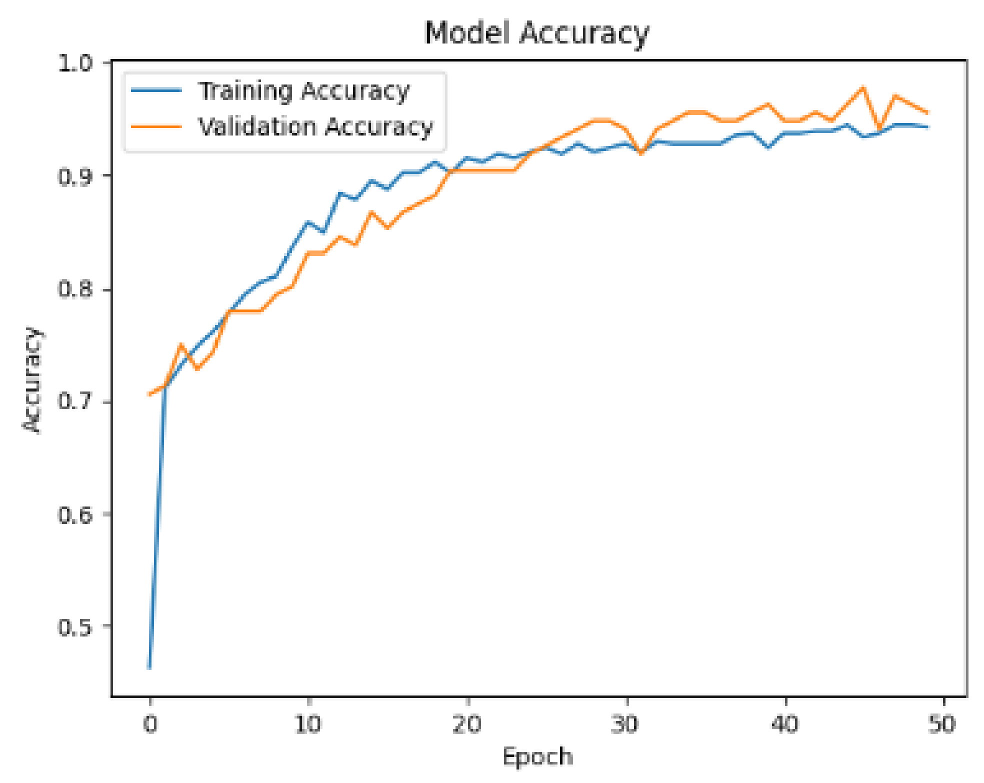
For class labels we used the same three different **importance functions:**

- Importance function based on normalized frequency
- Manually curated topics
- Importance based on cosine similarity

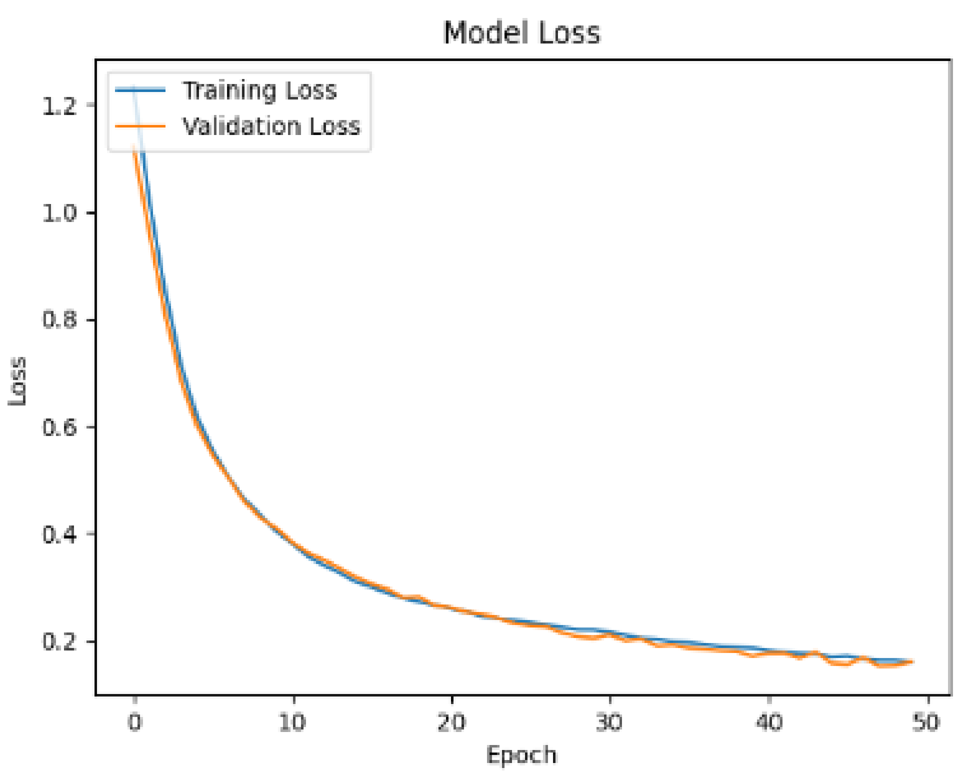
Neural Networks: Normalized frequency

The resulting classificaton report:

Topic	Precision	Recall	F1-Score	Support
NMF_topic_1	0.93	0.98	0.95	42
NMF_topic_2	0.98	0.95	0.97	60
NMF_topic_3	1.00	1.00	1.00	68
Accuracy		0.98		170
Macro Avg	0.97	0.98	0.97	170
Weighted Avg	0.98	0.98	0.98	170



(a) Training and validation accuracy values



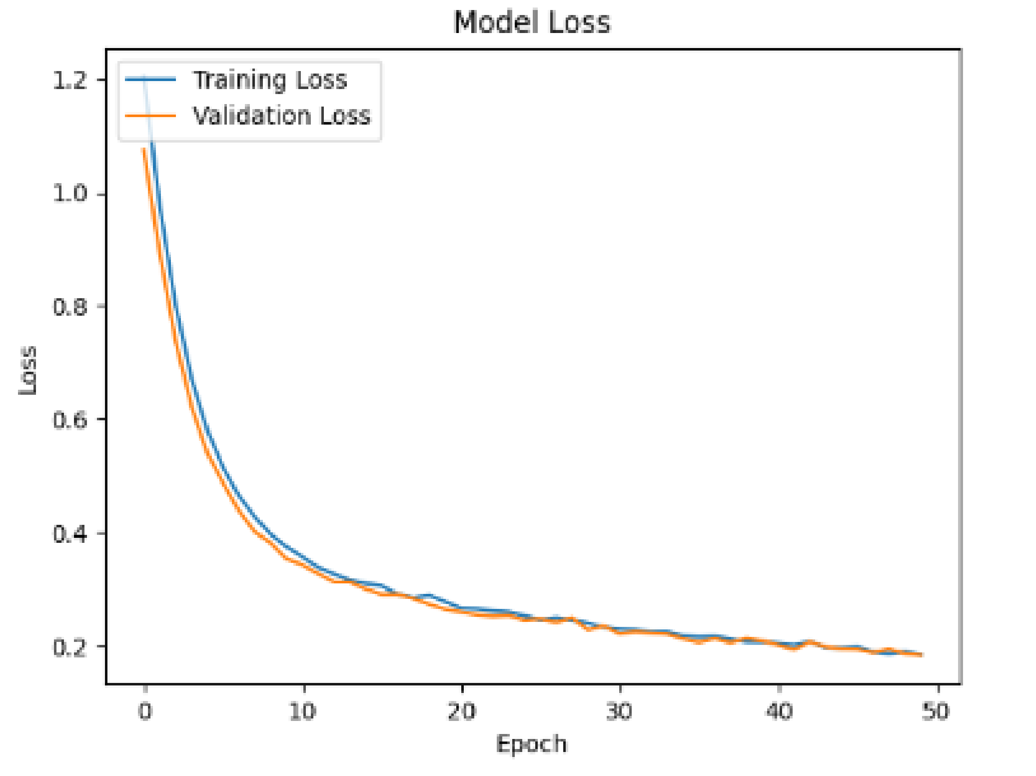
(b) Training and validation loss values

Neural Networks: Manually-curated topics

The resulting classificaton report:

Topic	Precision	Recall	F1-Score	Support
Topic_2	0.98	0.92	0.95	51
Topic_3	0.84	0.91	0.88	23
Topic_4	0.65	0.76	0.70	17
Topic_5	0.96	0.94	0.95	79
Accuracy		0.91		170
Macro Avg	0.86	0.88	0.87	170
Weighted Avg	0.92	0.91	0.91	170

(a) Training and validation accuracy values

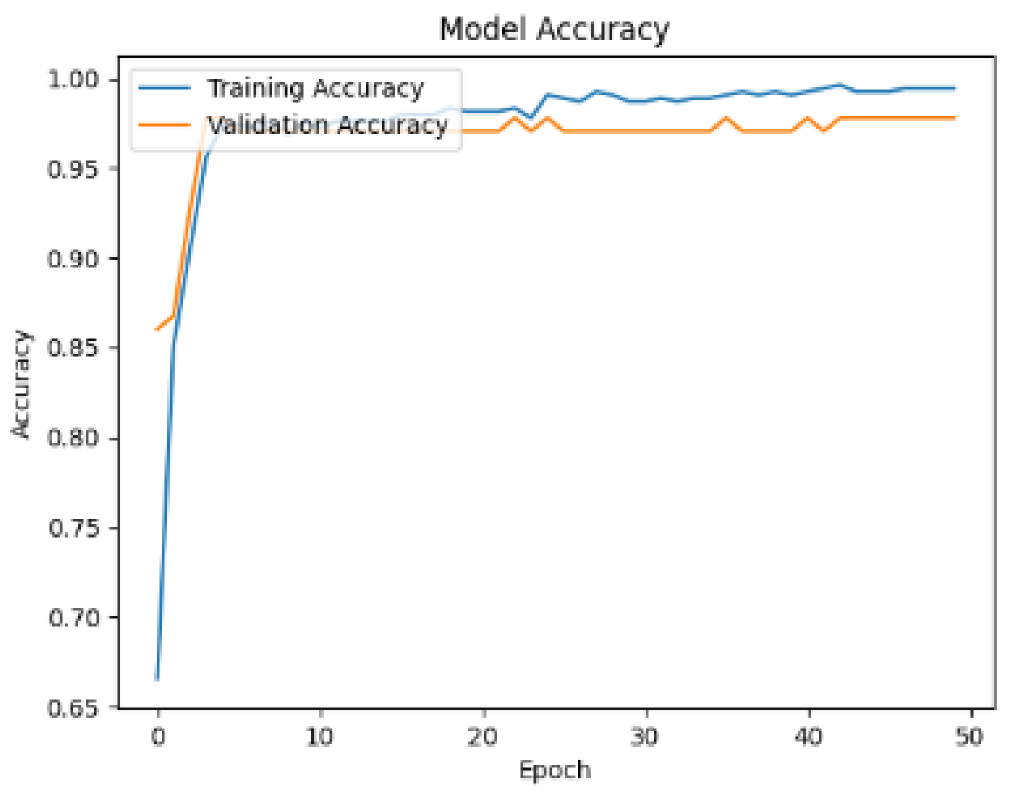


(b) Training and validation loss values

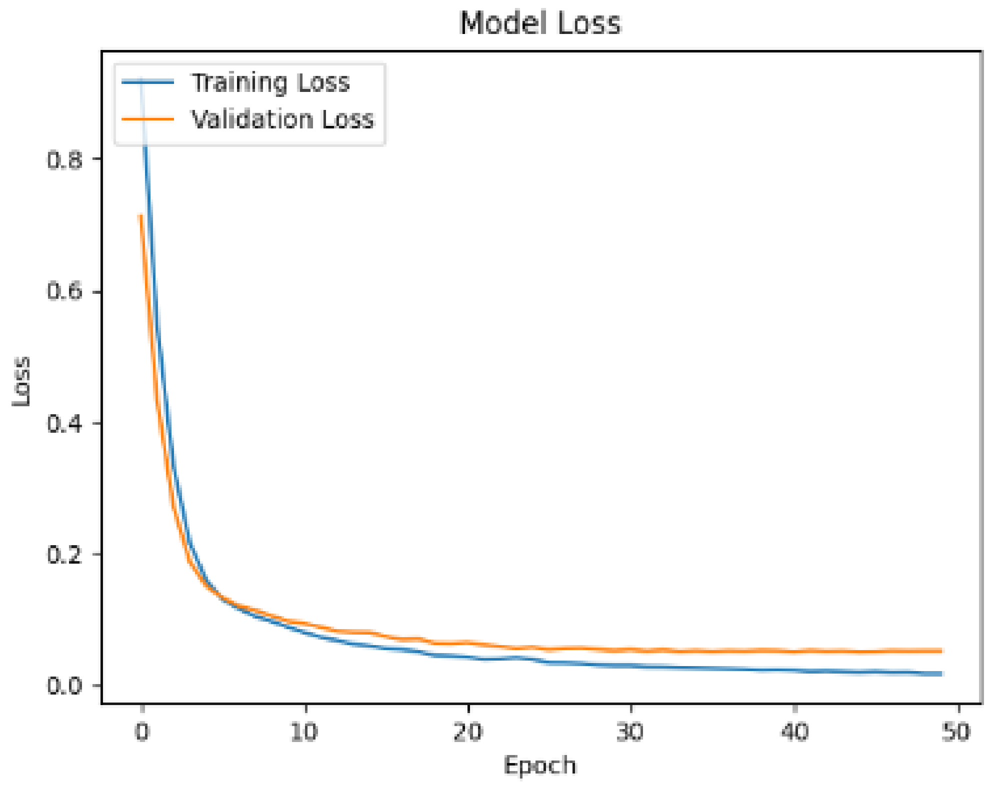
Neural Networks: Cosine similarity

The resulting classificaton report:

Topic	Precision	Recall	F1-Score	Support
Climate Risks to Heritage Sites	0.98	0.97	0.98	129
Flood Hazard and Vulnerability Assessment	1.00	0.80	0.89	5
Indoor Environmental Quality in Museums	0.90	0.97	0.93	36
Accuracy		0.96		170
Macro Avg	0.96	0.91	0.93	170
Weighted Avg	0.97	0.96	0.96	170



(a) Training and validation accuracy values



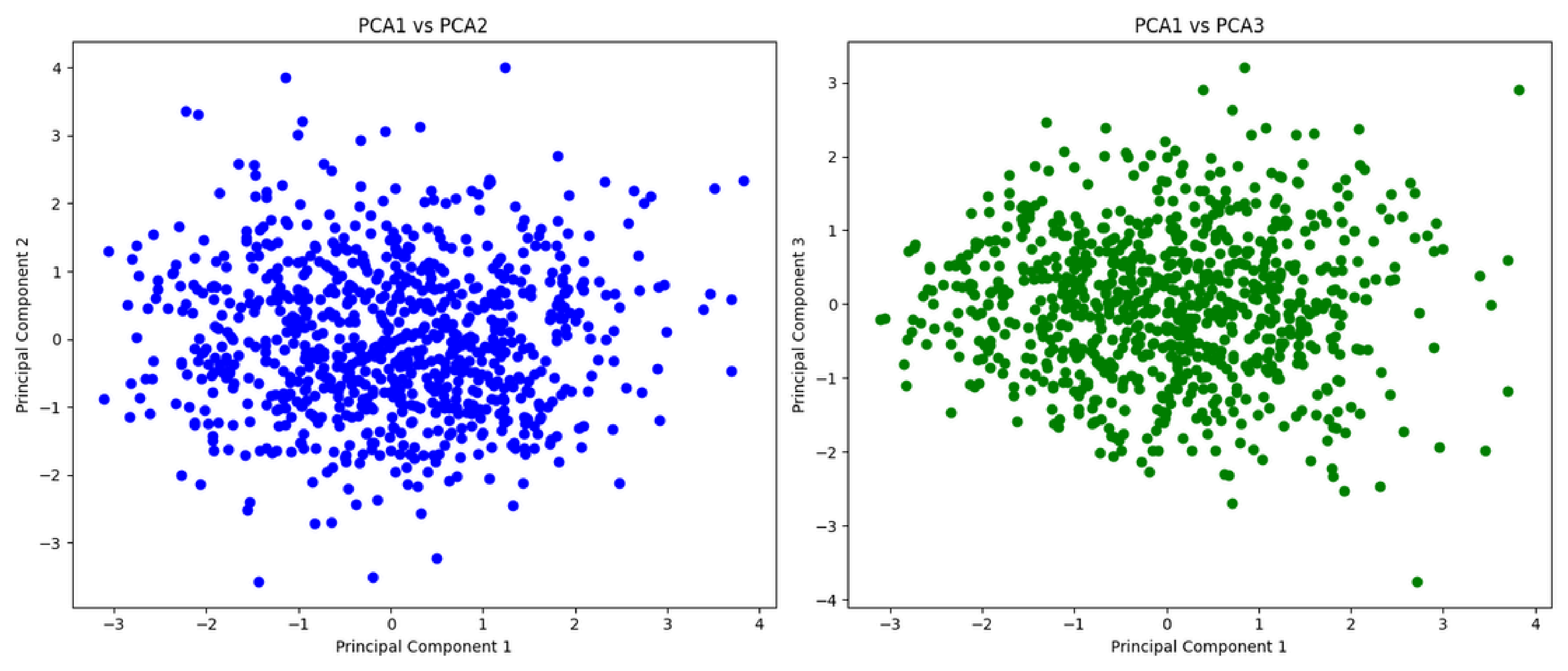
(b) Training and validation loss values

Bert

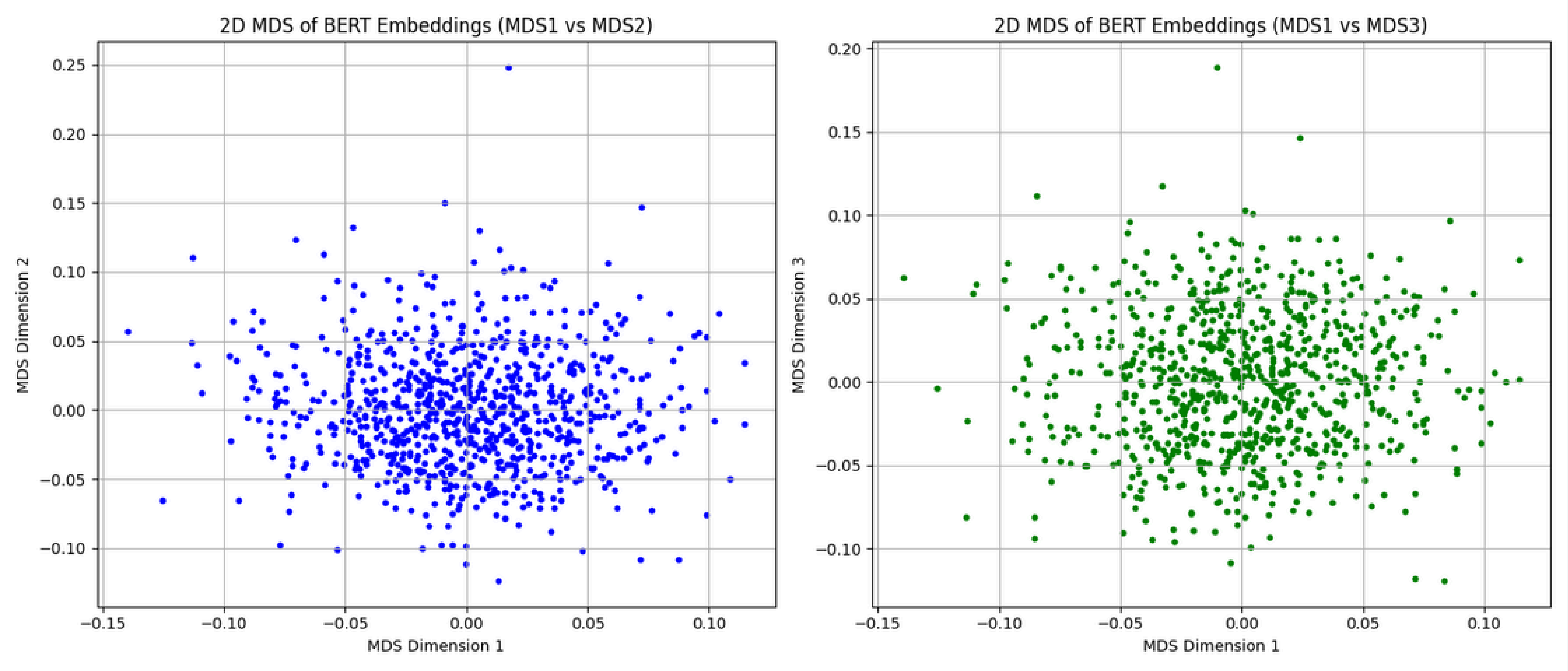
BERT is part of the NLP family and it employs a **bidirectional attention mechanism** to capture contextual relationships between words, allowing it to understand the meaning of a word based on its surrounding context. This results in richer representations of words, sentences, and documents.

We perform **PCA** and **MDS** on BERT embeddings, which allows us to reduce the high-dimensional space into a 2D or 3D visualization, making it easier to explore potential clustering patterns:

PCA



MDS



Bert: Random Forest models

We ran **3 different random forest** model with the aforementioned class labeling techniques, combining Bert embeddings with the probabilities recovered from the six different topic model techniques

Cosine similarity

Class	Precision	Recall	F1-Score	Support
Climate Risks to Heritage Sites	0.96	0.99	0.97	129
Flood Hazard and Vulnerability Assessment	0.00	0.00	0.00	5
Indoor Environmental Quality in Museums	0.97	0.97	0.97	36
Accuracy		0.96		170
Macro Avg	0.64	0.65	0.65	170
Weighted Avg	0.93	0.96	0.94	170

Manually curated topics

Class	Precision	Recall	F1-Score	Support
Topic_2	0.98	0.96	0.97	51
Topic_3	0.92	0.96	0.94	23
Topic_4	0.86	0.71	0.77	17
Topic_5	0.95	0.99	0.97	79
Accuracy		0.95		170
Macro Avg	0.93	0.90	0.91	170
Weighted Avg	0.95	0.95	0.95	170

Normalized frequency

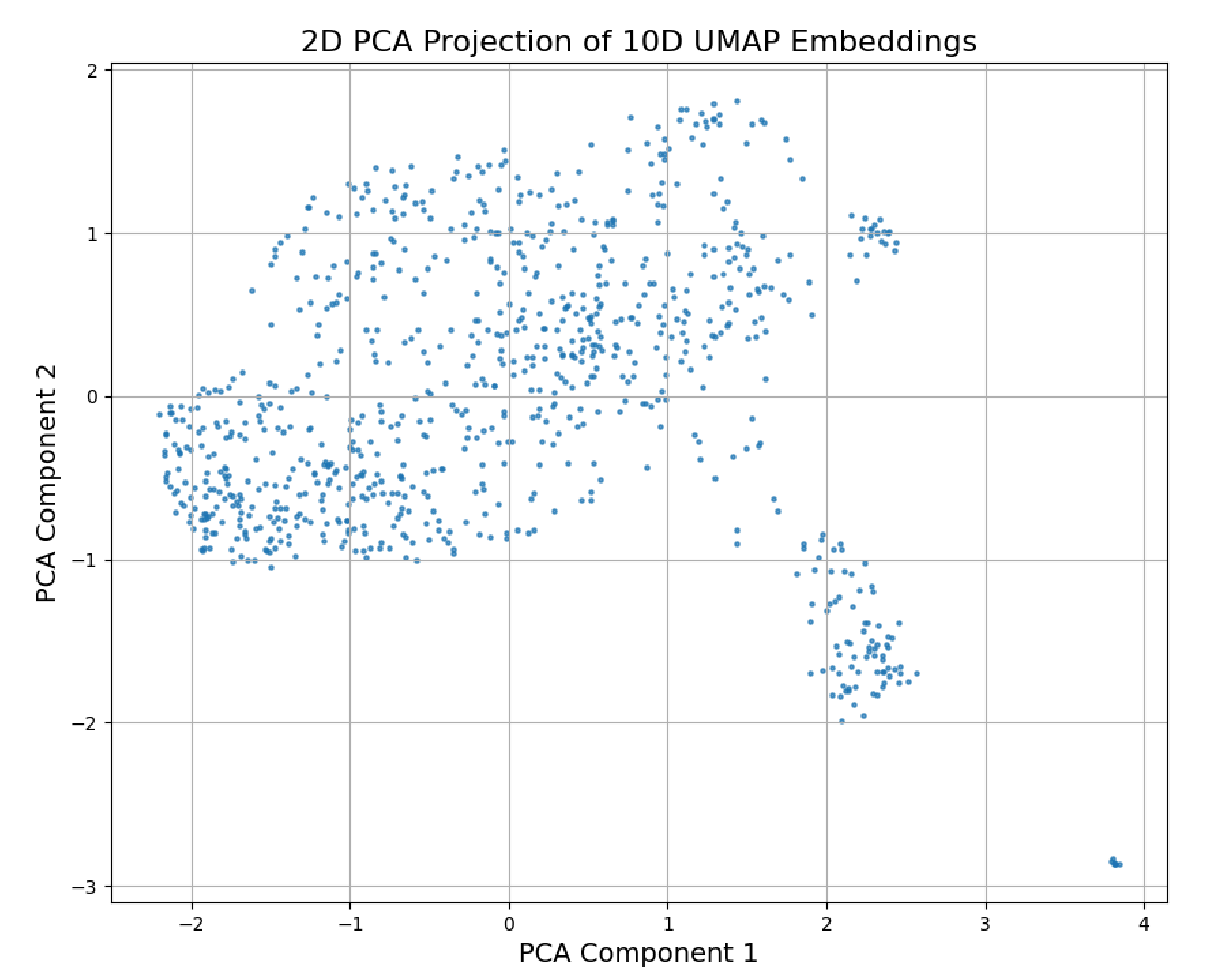
Class	Precision	Recall	F1-Score	Support
NMF_topic_1	0.95	1.00	0.98	42
NMF_topic_2	1.00	0.93	0.97	60
NMF_topic_3	0.97	1.00	0.99	68
Accuracy		0.98		170
Macro Avg	0.98	0.98	0.98	170
Weighted Avg	0.98	0.98	0.98	170

UMAP

Given the limitations of PCA and MDS in visualizing embeddings, we apply **Uniform Manifold Approximation and Projection (UMAP)** as an alternative dimensionality reduction technique.

We used parameters:

- **n_neighbors=15**: this parameter controls the balance between local and global structure.
- **n_components=10**: this parameter controls the target dimensionality of the embedding.
- **metric='cosine'**: we used the cosine distance metric, which is the choice for text data



LL Models: BERTopic

BERTopic is a **topic modeling technique** that uses BERT embeddings, UMAP for dimensionality reduction, and dynamic clustering to extract coherent topics, offering improved interpretability compared to traditional topic model techniques like LDA.

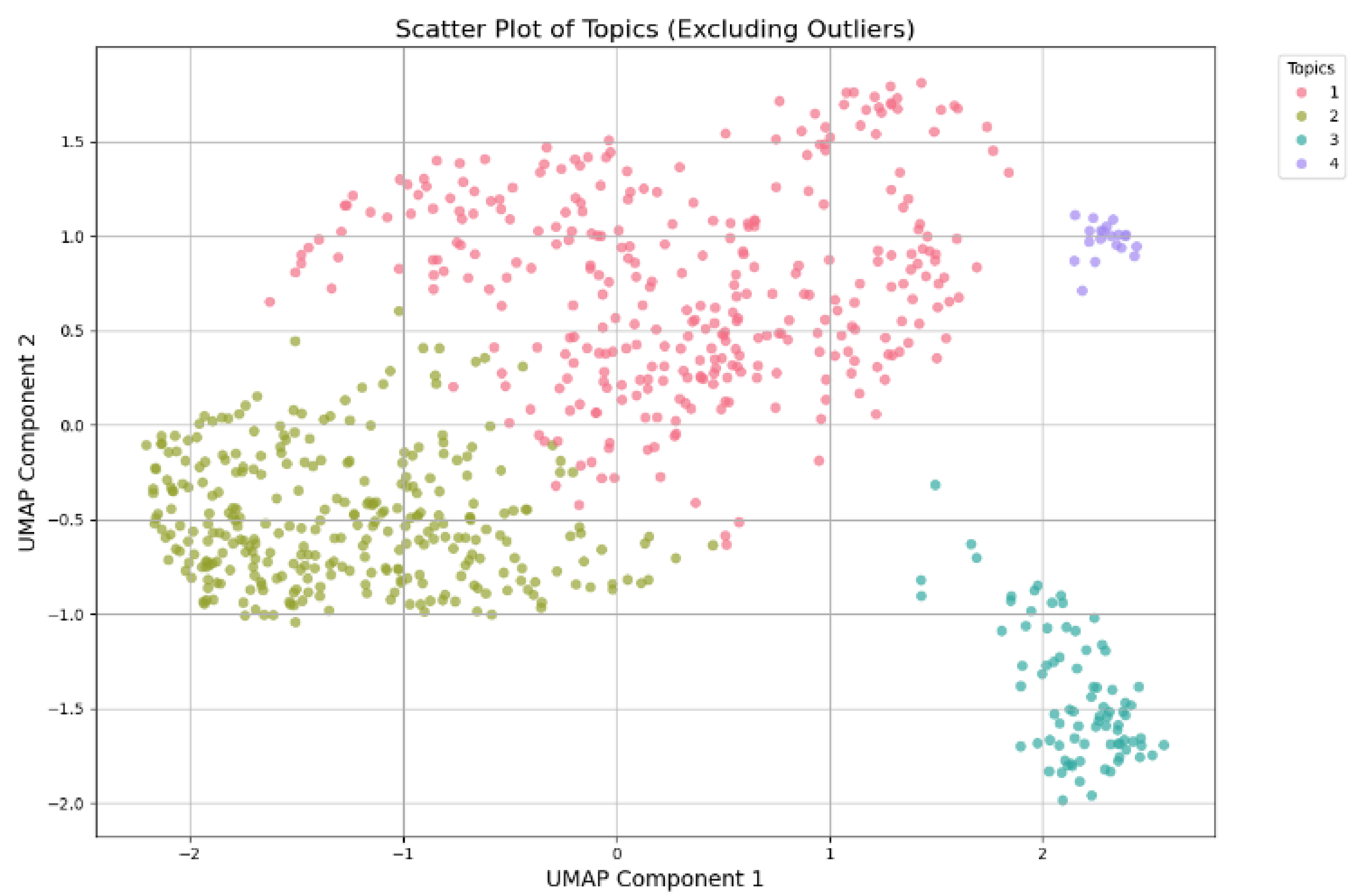
Identified Clusters:

1. **Climate Change's Impact on Cultural Heritage**
   Keywords: [heritage, climate, change, cultural, building, study, site, conservation, impact, environmental]
2. **Adaptation and Management Strategies**
   Keywords: [climate, change, heritage, cultural, site, impact, adaptation, management, risk, area]
3. **Flood Risk in Coastal Areas**
   Keywords: [site, heritage, risk, flood, hazard, cultural, coastal, area, study, change]
4. **Air Pollution Impact on Cultural Heritage**
   Keywords: [air, pollution, material, heritage, stone, surface, cultural, pollutant, study, building]

BERTopics

Three coherence scores assess topic quality:

- **c_v (0.5012)** combines statistical and semantic measures for balanced evaluation,
- **u_mass (-1.0361)** focuses on word co-occurrence probabilities with negative values indicating low coherence,
- **c_npmi (0.0651)** normalizes mutual information to measure word association strength



Class	Precision	Recall	F1-Score	Support
-1	1.00	0.56	0.72	16
0	0.95	0.98	0.96	112
1	0.96	1.00	0.98	100
2	1.00	0.93	0.97	15
3	1.00	1.00	1.00	12
Accuracy		0.96		255
Macro Avg	0.98	0.90	0.93	255
Weighted Avg	0.96	0.96	0.96	255

Clustering techniques applied to UMAP embeddings

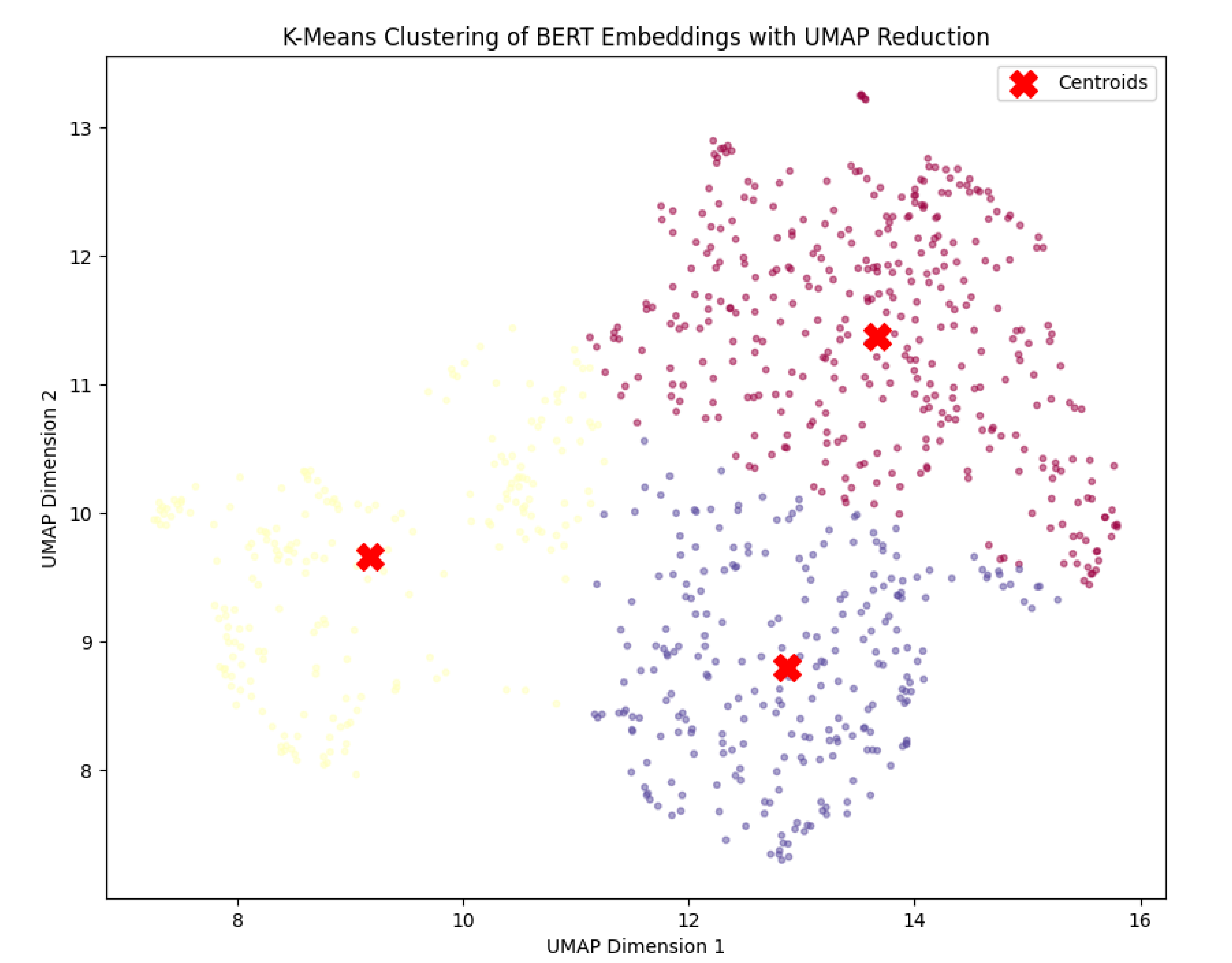
We further decided to apply **clustering techniques** to the UMAP embeddings to try to detect clusters in our data, we evaluated each technique with different metrics and selected K-Means as the best one:

Clustering Technique	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
HDBSCAN	0.0908	1.3545	153.46
K-means	0.3870	0.9327	621.35
Gaussian Mixture Model	0.0614	2.2372	152.67
OPTICS	-0.2702	1.0782	29.43
Spectral Clustering	0.2705	1.2130	400.19
Agglomerative Clustering	0.3333	0.9054	467.59

K-Means

We selected k=3 cluster given from the Elbow and Silhouettes analysis. Based on the analysis of the top unigrams and bigrams:

- **Cluster 1: Risk Management and Sustainable Development**
- **Cluster 2: Air Quality and Material Conservation**
- **Cluster 3: Flood Risks and Natural Hazards**



Class	Precision	Recall	F1-Score	Support
Class 0	1.00	0.98	0.99	81
Class 1	0.97	1.00	0.99	77
Class 2	1.00	1.00	1.00	12
Accuracy		0.99		170
Macro Avg	0.99	0.99	0.99	170
Weighted Avg	0.99	0.99	0.99	170

Ensemble Methods

We implemented two kinds of Ensemble methods:

- **Voting**
- **Stacking**

We will have 3 Ensemble models based on the **three different class labeling procedures:**

- Importance function based on normalized frequency
- Manually curated topics as class labels
- Importance function based on cosine similarity

Each **Ensemble Model** will comprehend three models:

- Random Forest model based on topic modeling (TM) probabilities,
- Random Forest model that combines BERT embeddings with TM probabilities,
- Feedforward Neural Network (FNN) model.

Voting - Summary

	Meta classifier accuracy	Meta classifier precision	Meta classifier recall	Meta classifier F1
Importance function based on Normalized frequency	0.9824	0.9825	0.9824	0.9824
Manually curated topics as class labels	0.9412	0.9396	0.9412	0.9399
Importance function based on Cosine Similarity	0.9765	0.9768	0.9765	0.9763

Stacking - Summary

	Meta classifier accuracy	Meta classifier precision	Meta classifier recall	Meta classifier F1
Importance function based on Normalized frequency	0.9824	0.9825	0.9824	0.9824
Manually curated topics as class labels	0.9412	0.9419	0.9423	0.9394
Importance function based on Cosine Similarity	0.9765	0.9768	0.9765	0.9763

Summary takeaways from classification

We have seen many different types of models and techniques in order to build the optimal classifier. The **main takeaways** are:

- The topics are very much overlapping
- The differences between each subtopic are subtle and too little to be detected
- Each document shares multiple topics and cannot be categorized under a single topic
- The real underlying topic is "Climate change's impact on cultural heritage"

From the previous analysis I can say that the main underlying topics in which we can actually cluster the dataset related to climate change are:

- **Flood risk**
- **Air pollution**
- **Management and assessment of risks**

Application to more papers

We aim to evaluate the classifiers developed on a **new dataset.** The initial dataset comprised papers published between 2016 and April 2024, sourced from the Web of Science and Scopus databases.

We conducted a new search on papers published between April 2024 and October 2024:

- Scopus: **256**
- Web of Science: **109**
- Total after duplicates removal and abstract review: **259**

Each abstract was manually annotated:

- **Topic 1:** Climate Change's Impact on Cultural Heritage
- **Topic 2:** Adaptation and Management Strategies
- **Topic 3:** Flood Risk and Water-Related Damages in Coastal Areas
- **Topic 4:** Air Pollution Impact on Cultural Heritage

Application of the models

We applied the following classifiers and predicted the labels on the new dataset:

- ~~K-means Bert classifier~~
- BerTopics classifier
- Random Forest model based on normalized frequency
- Random Forest model with manually annotated topics
- ~~Random Forest model based on cosine similarity~~
- Random Forest model based on normalized frequency combined with BERT embeddings
- Neural Network model based on normalized frequency
- Neural Network model with manually annotated topics
- Neural Network model based on cosine similarity

We decided to **exclude** from downstream analysis the **K-means Bert classifier** and the **Random Forest model based on cosine similarity** as they predicted the same label for every document in the dataset.

Label Mapping

We **standardized the naming conventions** for consistency. We also chose to merge Topics 1 and 2 due to their semantic overlap and frequent interchangeability.

- **Topic 1**= "Climate Change's Impact on Cultural Heritage and adaptation and management strategies",
- **Topic 2** = "Flood Risk and Water-Related Damages in Coastal Areas",
- **Topic 3** = "Air Pollution Impact on Cultural Heritage Surface Quality".

Topic	Normalized Frequency	Manually annotated	Manually curated	Cosine	Bertopics
Topic 1	NMF_topic_3	Topic_1, Topic_2	Topic_3, Topic_5	0	Topic_1, Topic_2
Topic 2	NMF_topic_2	Topic_3	Topic_4	2	Topic_3
Topic 3	NMF_topic_1	Topic_4	Topic_2	1	Topic_4

Performance

To evaluate the performance of each model's predictions, we compared them against the topics that we manually annotated.

Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa *
Manually Curated Renamed RF	0.710425	0.655760	0.710425	0.677574	0.372233
Bertopics Renamed	0.555985	0.448737	0.555985	0.493049	-0.077488
Manually Curated Renamed NN	0.471042	0.632839	0.471042	0.486656	0.205735
Cosine Renamed NN	0.459459	0.520525	0.459459	0.478432	0.033092
Norm Frequency Renamed NN	0.378378	0.573806	0.378378	0.411663	0.086910
Norm Frequency Renamed RF + BERT	0.362934	0.675019	0.362934	0.394215	0.068609
Renamed RF Norm Frequency	0.362934	0.440580	0.362934	0.370623	-0.044125

***Cohen's Kappa** is a statistical measure used to assess the agreement between two raters or classification models. It accounts for the agreement that could occur by chance, providing a more robust measure than simple accuracy.

Performance takeaways

Method reviewed:

- **Best Method:** Manually_curated_renamed_rf
- **Moderate Performer:** Bertopics_renamed
- **Weakest Methods:** Norm_freq_renamed_rf_bert and Renamed_rf_norm_freq

Overall Insights:

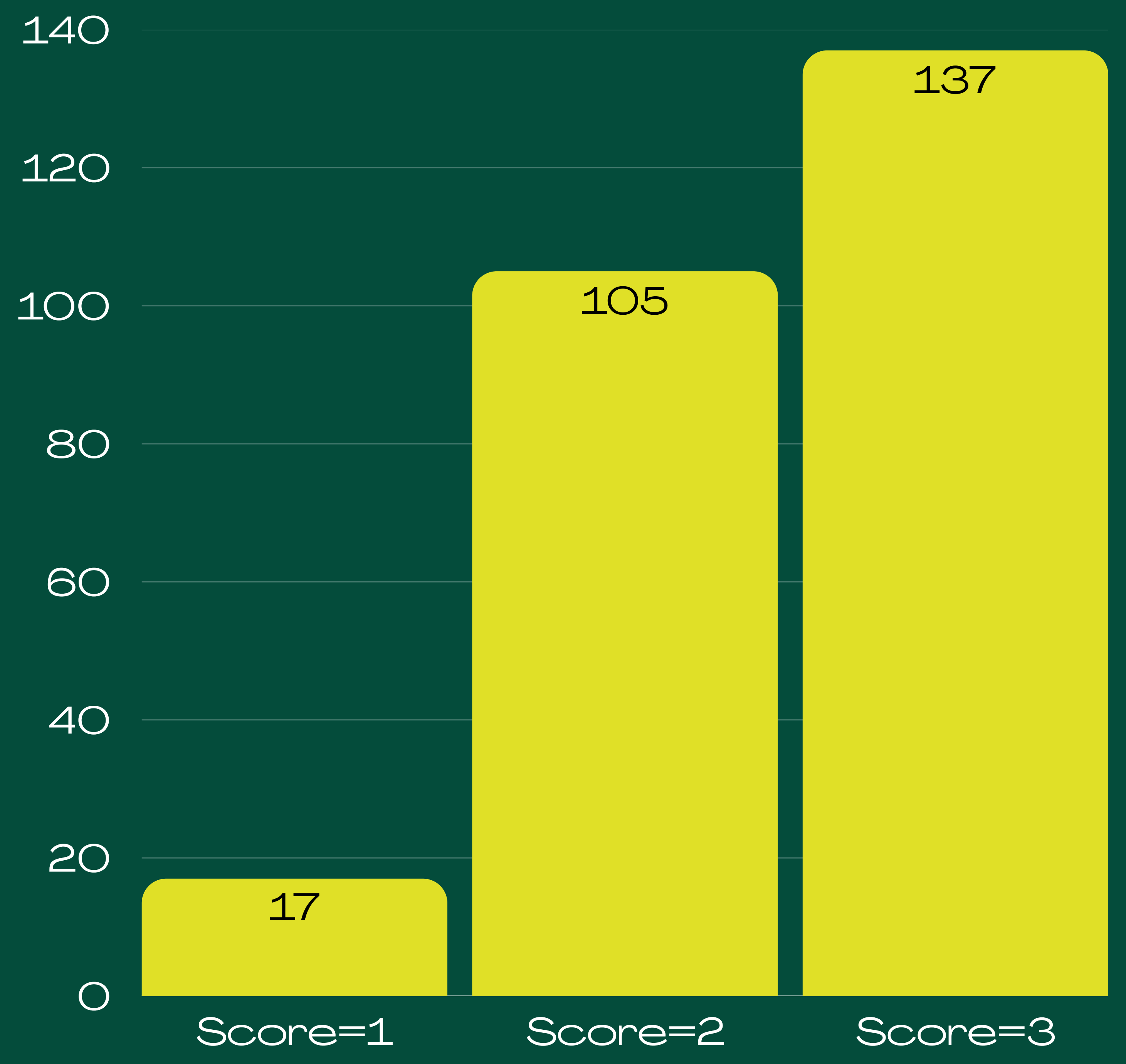
- Top Method: Manually_curated_renamed_rf shows reliable classification ability.
- Other Methods: Require optimization, particularly neural networks.
- Subjectivity in topic assignment; overlap complicates dominant topic designation.
- Topic assignments derived from abstracts, not full paper content.

Agreement Score

A function was implemented to calculate an **Agreement Score** for each document, measuring consensus among model predictions. The score counts unique predictions:

- **A lower score** indicates higher agreement (1 signifies unanimous predictions)
- **A higher score** indicates greater disagreement

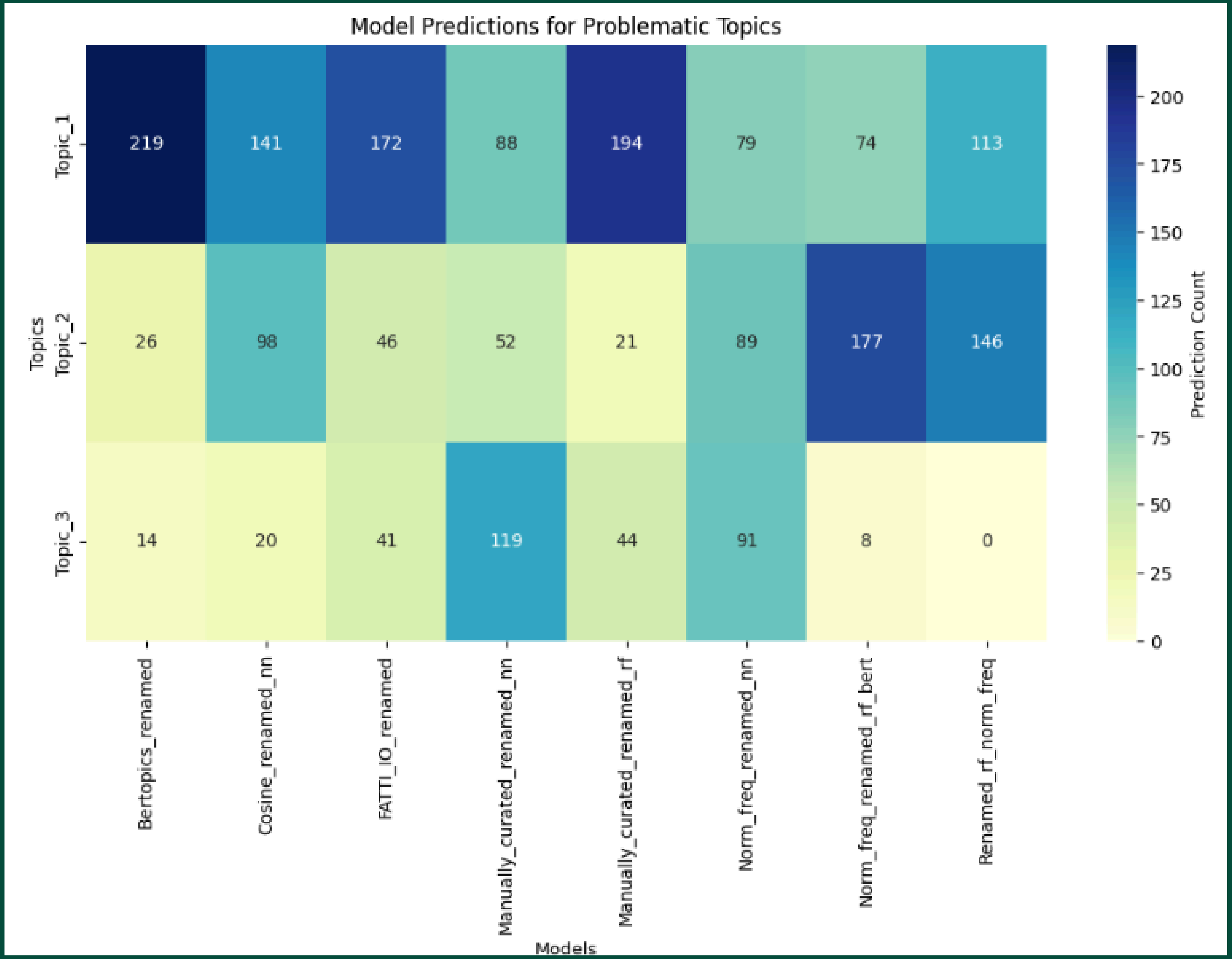
Out of 259 documents:



Agreement between predictions

We developed a function to evaluate and visualize predictions from various models for our topics, focusing on areas of agreement and disagreement.

- **Topic_1:**
  - **Highest Agreement:** Random Forest models and manually annotated topics
  - **Moderate Counts:** Neural Networks and Bertopics
  - **Lower Counts:** Some Neural Network
- **Topic_2:**
  - **Highest Counts:** Bertopics and Neural Networks
  - **Mid-Level Counts:** Other Neural Networks
  - **Lower Counts**: Random Forest models
- **Topic_3:**
  - **Highest Counts:** Neural Networks and manually annotated topics
  - **Low/Zero Counts:** Several models



Heatmap takeaways

Heatmap Insights on Model Agreement and Divergence:

- **Model Agreement:** Topic_1 shows the highest agreement across models (e.g., Random Forest + BERT), likely due to its clear distinguishing features.
- **Topic Distinguishability:** Topic_3 has the lowest prediction rates, indicating overlap with other topics or difficulty in clear definition.
- **Model-Specific Behavior:** Models like BerTopics and Neural Networks (manual topics) favor Topic_2, while others (e.g., Neural Network on normalized frequency) show more balanced predictions due to differing criteria.

Key Takeaways on Topic Modeling and Classification

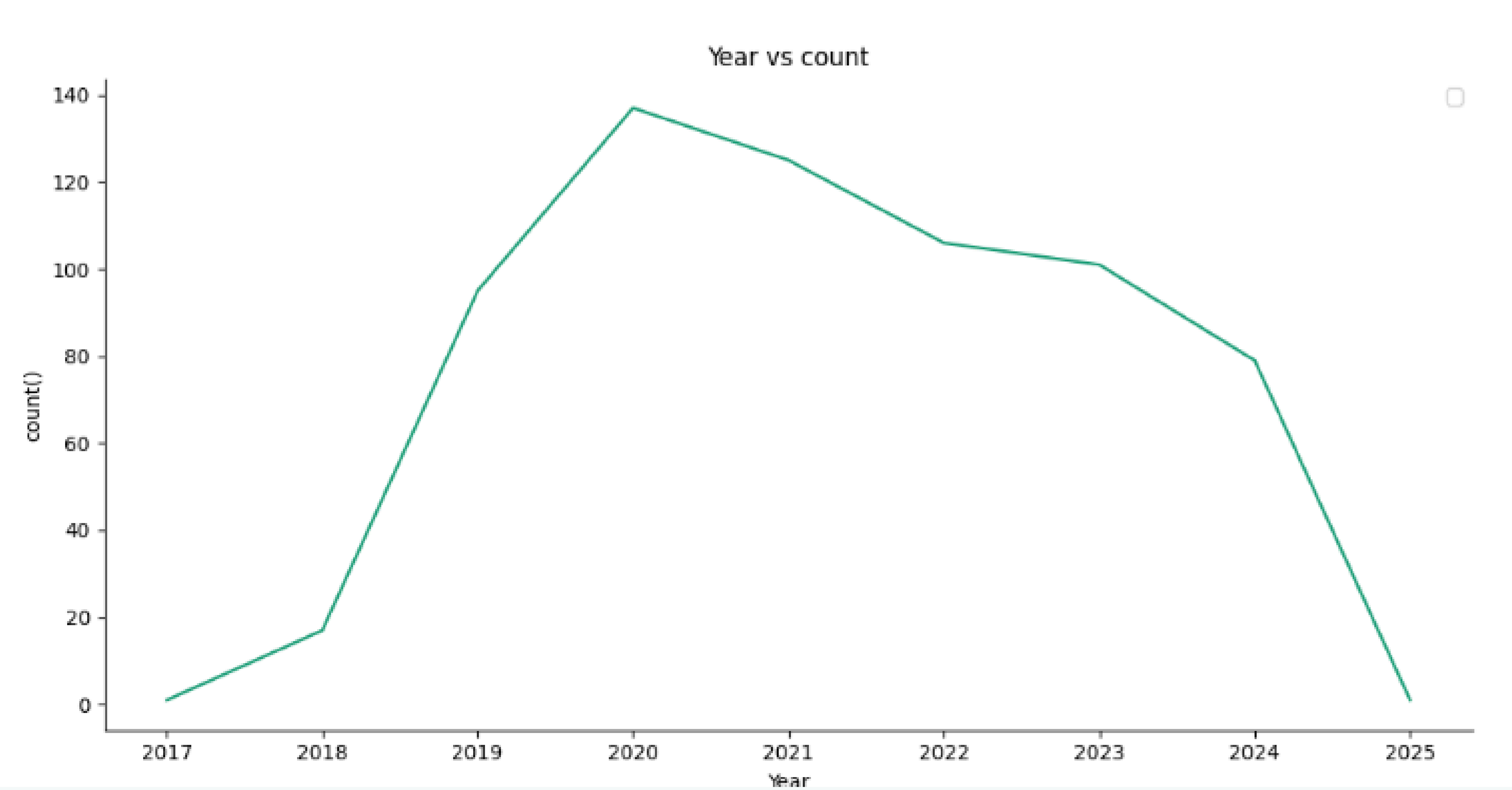
- **Topic Overlap:** Many topics are distinct but frequently overlap due to hierarchical relationships (major topics vs. subtopics).
- **Multi-Label Complexity:** The problem requires multi-label classification to address overlapping and hierarchical themes.
- **Data Scope:** Analyzing full papers, not just abstracts, could provide deeper insights.
- **Topic Models' Utility:** Topic modeling is a valuable technique for identifying main themes in literature.
- **Classification Challenges:** Most classifiers misclassify new documents under "major" topics due to overlapping features.
- **Best Technique:** NMF emerged as the most effective topic model, providing the best class labels for the top-performing classifier.

Overtourism

Overtourism is a pressing global phenomenon characterized by **excessive tourist numbers** in popular destinations, leading to negative impacts on local communities, environments, and cultural heritage.

We conducted a systematic search using the keyword "overtourism":

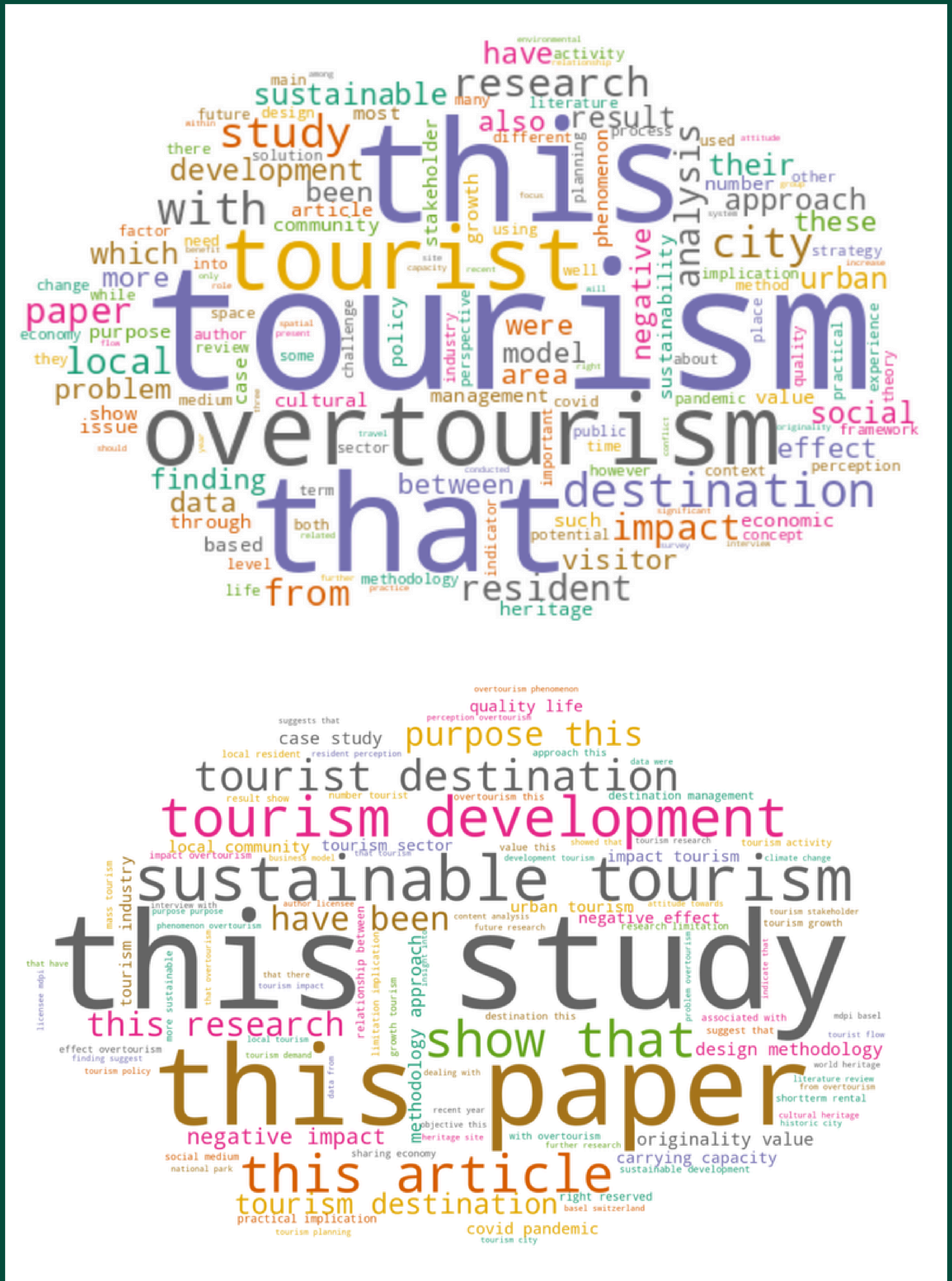
- Scopus: 439 articles
- Web of Science: 471 articles
- After removing duplicates: 698 articles.
- Further refinement: **662 articles** suitable for analysis.



Preliminary analysis

These are the most interesting words that emerged from **unigrams** and **bigrams,** reflecting core themes in the dataset as:

- tourism,
- destination,
- heritage,
- negative impact,
- sustainable tourism,
- tourism development,
- community,
- quality of life,
- local resident



TM results

Technique	N° of Topics
LDA	8
LSI	6
HDP	8
NMF	10
STM	6
CTM	6

Most common topics	Techniques
Overtourism and its impact	LSI, HDP, STM, CTM
Cultural Heritage and Identity	LSI, HDP, NMF, STM
Sustainable Tourism Practices	LDA, LSI, HDP, NMF, STM CTM
Urban Dynamics and Challenges	LDA, LSI, HDP, STM, CTM
Community Perspective and Resident Impact	LSI, HDP, NMF, CTM
Economic Factors and Development	LSI, STM, HDP, NMF
Visitor Experience and Engagement	LDA, LSI, NMF, STM, CTM

Key Takeaways on Overtourism

- **Multifaceted Impact:** Overtourism affects cultural heritage, community dynamics, and sustainable tourism practices.
- **Interdisciplinary Nature:** Tourism research integrates themes like sustainability, cultural impact, and community perspectives, requiring holistic management approaches.
- **Policy Needs:** Effective strategies demand informed policymaking and governance that consider all stakeholders (local communities, businesses, and visitors).
- **Future Research Opportunities:** Explore innovative solutions for sustainable tourism, stakeholder engagement, and community resilience.

Next steps

Climate Change’s Impact on Cultural Heritage

- Enhance existing models to improve their accuracy and applicability.
- Explore a broader range of models, with a particular focus on leveraging advancements in large language models (LLMs).
- Continuously integrate new and relevant literature as it becomes available.
- Conduct analyses using the full content of research papers rather than relying solely on abstracts.

Overtourism

- Expand the inclusion of relevant literature to deepen insights.
- Develop and refine classification models to better categorize and analyze trends.
- Perform comprehensive analyses based on the full text of research papers instead of limiting to abstracts.

Grazie per l'attenzione