Master Thesis

Trend and Popularity Analysis for Art Related Texts from 1600-1800

Tessel Haagen

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Antske Fokkens
2^{nd} reader: Isa Maks

Submitted: August, 2025

Abstract

This thesis investigates evolving perceptions of nature in early modern English texts by combining topic modeling, sentiment analysis, and trend detection. Drawing from the EEBO and ECCO corpora, it analyzes paragraphs containing art-related keywords while excluding highly religious and theater texts. The primary methodological framework applies BERTopic with MacBERTh embeddings to identify latent themes, complemented by sentiment analysis using TextBlob and visualized trends over time. A secondary framework uses GloVe word embeddings to expand keyword searches for targeted topic analysis. Evaluations of the frameworks indicate potential, as several topics exhibit interpretability, and their frequency are aligned with established trends in art historical knowledge. However, it is crucial to address limitations pertaining to semantic and topic alignment to fully leverage these frameworks.

Declaration of Authorship

I, Tessel Eva Haagen, declare that this thesis, titled *Trend and Popularity Analysis for Art Related Texts from 1600-1800* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I used AI as a writing assistent tool for the abstract, the discussion and the conclusion chapters.

Date: 18-08-2025

Signed: 

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Art historians have long relied on textual evidence to understand art's cultural, social, and intellectual dimensions (Rabb and Brown, 1986). Texts such as treatises, critiques, journals, and other written records provide invaluable insights into how artistic movements, styles, and ideas were perceived, discussed, and valued over time. These sources document evolving artistic practices and reveal the shifting terminologies and concepts that shaped the reception of art. However, identifying and researching these patterns presents significant challenges. Recognizing meaningful trends requires an initial hypothesis, and verifying them demands extensive manual investigation across large, diverse texts. For instance, tracing the origins of a term—determining when and where it first appeared, how its meaning evolved, and in what contexts it was used—requires systematically analyzing vast amounts of historical documents. Similarly, assessing the frequency of specific terms over time to identify shifts in artistic discourse is difficult without computational assistance, as it involves locating and quantifying every instance of usage across different sources and periods. Manually conducting such analyses is not only time-consuming but also prone to inconsistencies and human error, making it difficult to uncover broader trends without structured, automated methods.

Advancements in text mining and natural language processing (NLP) offer a transformative solution to these challenges. The digitization of historical texts has made it possible to analyze vast amounts of written material computationally, enabling new ways to study artistic discourse. By leveraging computational tools, researchers can uncover trends and patterns in textual data that would otherwise remain undetected (Bowman, 2023; Chen et al., 2022). NLP techniques help automate and enhance these processes. Word embeddings, for instance, allow researchers to study the semantic relationships between words, revealing how the meanings of artistic terms shift over time (Kozlowski et al., 2018). Topic modeling groups related words together, identifying overarching themes within large text collections and highlighting changes in artistic discourse (Roose et al., 2018). Sentiment analysis can assess how certain artistic styles or movements were received, capturing shifts in perception through emotional tone. These techniques provide a more systematic, scalable, and efficient approach to textual analysis, making it possible to identify key terms, rising domains, and latent topics within large corpora in ways that were previously impractical.

The changing popularity of landscape painting in England provides an interesting example of how shifts in artistic discourse can reflect broader cultural trends. Over time, landscape paintings in England evolved from simple backgrounds to significant symbolic representations of national pride (Budnick, 2017). Identifying and analyzing

the terms and sentiments associated with landscape painting during this transformation requires an effective approach to handling large volumes of textual data. For example, the increasing prominence of terms like "picturesque" in landscape painting writing signals the growing significance of the genre and the emergence of new aesthetic ideals (Townsend, 1997). This thesis aims to meet the needs of art historians by developing and implementing two alternative text-mining frameworks. The first framework is intended for general trend analysis, and the second framework concentrates on specialized trend analysis, like the use case described here.

1.1 Research Goal, Relevance, and Contribution

This research seeks to contribute to the ongoing efforts to bridge the gap between computational linguistics and traditional humanities scholarship. Conventional approaches are still quite useful because of their variety of interpretation and depth, but they are frequently time-consuming and have a narrow focus. Text mining, on the other hand, provides a scalable method that enables researchers to analyze enormous text datasets in a matter of seconds and uncover patterns that could otherwise go unnoticed. In addition to increasing productivity, this use of computational methods in art historical study offers an enhanced understanding of how language both reflects and shapes the cultural and intellectual reception of art.

The methodology of this thesis includes a multi-step approach, beginning with the collection of texts from Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO). Topic modeling will provide a thematic overview of the texts, while sentiment analysis will assess the perception and reception of artistic topics. The framework will be evaluated through a focused case study, examining how rising terms in artistic discourse can signal broader cultural and aesthetic trends.

This research contributes to the expanding field of digital humanities by providing a practical framework for the computational analysis of art historical texts. It emphasizes the potential of text-mining techniques to enhance traditional methodologies. Rather than aiming to prove the superiority of these tools, the study explores their potential, limitations, and relevance to the art historian's toolkit. By showcasing the benefits of computational approaches through a case study on the language of landscape painting, this thesis aims to encourage further interdisciplinary collaboration and innovation in the study of art history.

The findings indicate that topic modeling, which was carried out using BERTopic, had mixed results. It can successfully identify some clusters related to significant artistic and cultural themes, though some topics proved harder to interpret. A topic intrusion survey revealed that while many topics were coherent, others were challenging for annotators, pointing to the need for further refinement in preprocessing and modeling. Sentiment analysis showed a weak correlation with human evaluations, suggesting it can capture some emotional trends but needs more domain-specific calibration. Despite these challenges, the temporal analysis of themes offered valuable insights into how cultural priorities evolved over time. Overall, the results highlight the promise of digital methods to complement traditional approaches and open new directions for interdisciplinary art historical research.

1.2 Outline

First I will provide you with background on topic modeling and sentiment analysis, and a literature review of text-mining techniques used for humanities and art in Chapter 2. In Chapter 3, the methodology of this research will be explained. In Chapter 4 I will discuss the surveys conducted for the evaluation of the models. Next, I will discuss the results of the models, and the results for both frameworks, the general trends and the use case of landscape paintings in Chapter 5. Finally, I will go over the discussion points, and further work directions in Chapter 6 and conclude this study in Chapter 7.

Chapter 2

Literature and Background

This chapter presents an overview of topic modeling and sentiment analysis, two tasks that will be utilized in this thesis. I discuss various techniques for addressing these tasks and explore methods for evaluating the models. Finally, I delve into the challenges associated with using historical texts and review related work in this area.

2.1 Background of Topic Modeling

Topic modeling is a branch of machine learning and natural language processing (NLP) that focuses on uncovering latent topics within large sets of text data. Latent topics are a collection of words. Several methods have been developed over time, each employing different techniques for identifying hidden structure and semantics in the data.

2.1.1 Topic modeling techniques

In Figure 2.1, a hierarchy based on a couple of surveys of different topic model techniques and algorithms is shown (Murshed et al., 2023; Kherwa and Bansal, 2019). These techniques are covered in more detail below.

Probalistic methods

Probabilistic methods aim to capture the distribution of words over topics and topics over documents based on probabilistic frameworks.

One of the first probabilistic modeling methods is Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), which was a significant advancement in the field of topic modeling and information retrieval at the time. The approach works by assuming that there are hidden (latent) topics that generate the words observed in documents. It models each document as a mixture of these topics, and each topic as a probability distribution over words. The key idea is that the words in a document are generated by first choosing a topic according to the document's topic mixture, and then selecting a word from that topic's word distribution. The model starts with a document-term matrix, where each entry represents how often a term appears in a document. PLSA then uses this data to estimate two main probability distributions: the probability of a topic given a document, and the probability of a word given a topic. These distributions are estimated using the Expectation-Maximization (EM) algorithm, an iterative method that alternates between estimating the hidden variables (topic assignments) and updating the model parameters. PLSA offers several advantages, providing a principled

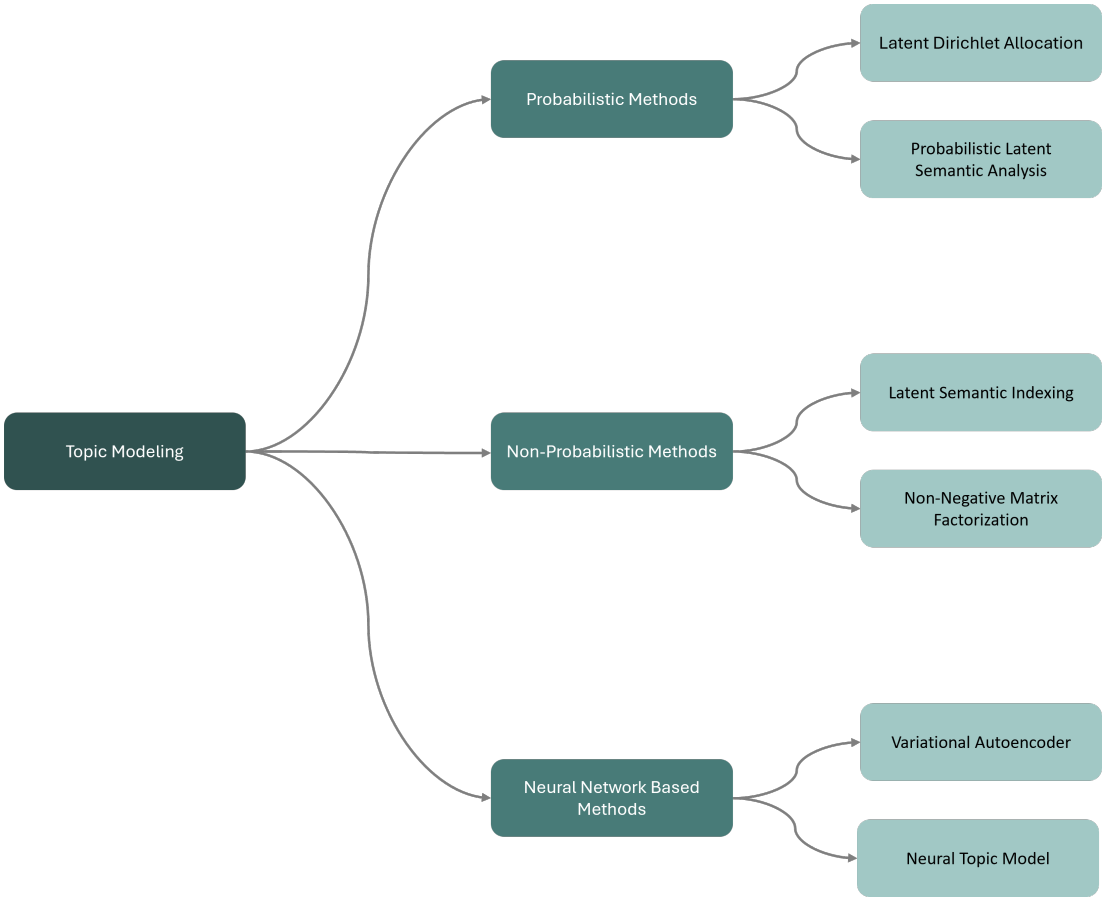


Figure 2.1: Hierarchy of Topic Modeling approaches.

probabilistic approach, which allows the handling of polysemy (words with multiple meanings) and flexible modeling of text data. However, PLSA also has limitations. It can be prone to overfitting, especially with large document collections, and it does not provide a generative model for new, unseen documents.

One of the most widely used probabilistic models for topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). It works under the same two key assumptions as PLSA: documents are composed of a mixture of topics, and topics are probability distributions over words. The algorithm assumes each document contains a small number of topics. Each word in a document is attributable to one of those topics. Then it uses word co-occurrences to identify sets of terms that likely belong to the same topic. It generates topic-word distributions: the probability of words appearing in each topic, and document-topic distributions: the proportion of each topic present in a document. The model treats documents as "bags of words", ignoring word order. It uses Bayesian inference and typically employs Gibbs sampling to assign words to topics. The algorithm requires specifying the number of topics beforehand. In the years after, the authors from Blei et al. (2003) researched extensions on LDA. Dynamic Topic Models (DTM) extend LDA to incorporate the temporal aspect of document collections (Blei and Lafferty, 2006). Correlated Topic Models (CTM) allow topics to be correlated rather than independent (Blei and Lafferty, 2007). By using a logistic normal distribution to model topic proportions, the approach allows topics to be correlated with one

another. Instead of assuming that topics are independent, the model uses a logistic normal distribution to represent topic proportions, which makes it possible to capture correlations between topics. Hierarchical LDA (hLDA) is a non-parametric extension of LDA (Blei et al., 2010). It automatically determines the number of topics by using a Dirichlet process (Teh et al., 2010).

Non-Probabilistic Methods

Non-probabilistic methods use linear algebra instead of probability theory to find topics.

Latent Semantic Indexing (LSI), also known as Latent Semantic Analysis (LSA), is one of the earliest topic modeling techniques, developed in the 1980s (Deerwester et al., 1990). LSI works by analyzing the patterns of word usage across a large set of documents. It assumes that words that frequently appear together in similar contexts are likely to have related meanings. This approach allows LSI to go beyond simple keyword matching and capture the conceptual content of documents. The core of LSI is a mathematical technique called Singular Value Decomposition (SVD). This process starts with a term-document matrix, where each row represents a unique word, and each column represents a document. The entries in this matrix typically represent the frequency of each word in each document. SVD then decomposes this large, sparse matrix into smaller, dense matrices that capture the most important patterns in the data. By reducing the dimensionality of the original term-document matrix, LSI creates a "semantic space" where both terms and documents are represented as vectors. In this space, semantically similar terms and documents are positioned close to each other. This allows LSI to identify relationships between terms that do not necessarily co-occur directly but are used in similar contexts across the document collection. One of the key strengths of LSI is its ability to handle synonymy (different words with similar meanings) and polysemy (words with multiple meanings). It has limitations as well; it does not handle negation well, and its effectiveness can decrease with very large or dynamic document collections.

Another method is Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999), which decomposes the document-term matrix into two non-negative matrices: a document-topic matrix and a topic-term matrix. This decomposition allows for the discovery of latent topics within a corpus of documents. NMF approaches topic modeling from a linear algebra perspective, offering a different set of strengths and interpretations. The non-negativity constraint in NMF leads to a parts-based representation of topics, where each topic is characterized by a combination of words, and each document is represented as a mixture of these topics. In practice, this approach involves setting the number of desired topics and iteratively refining the factorization to minimize the reconstruction error between the original document-term matrix and its approximation. This approach often results in more interpretable and coherent topics compared to other methods, however, it requires careful consideration of parameters and initialization methods to yield optimal results.

2.1.2 Text Representation Methods

Modern text representations, particularly word embeddings, have significantly enhanced topic modeling. Word embeddings, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2019), capture semantic and syntactic relationships between words in a continuous vector space, providing richer con-

textual information than traditional bag-of-words approaches, where the frequency of words was counted. These advancements have led to more nuanced and comprehensive text representations, facilitating better performance in various downstream NLP tasks. In topic modelling, various methods were proposed using these text representations. The Embedded Topic Model (ETM) combines these word embeddings with traditional topic modeling techniques, allowing for more interpretable topics even with large vocabularies that include rare words (Dieng et al., 2020). Another word embedding-enhanced topic model, WELDA (Word Embedding Latent Dirichlet Allocation), has demonstrated superior accuracy in news topic recognition tasks, reaching up to 97% accuracy on certain datasets (Kaleem et al., 2024). This integration enables topic models to leverage both local word collocation patterns and broader semantic relationships, resulting in improved topic quality and predictive performance (Zhang et al., 2019). Others disregard topic models and use clustering on the word embeddings to get topics (Sia et al., 2020).

Neural Network-Based Methods

The discovery of neural networks significantly advanced Neural Language Processing models, especially about ten years ago, thanks to the increased computational power. This also influenced topic modeling. Neural Topic Modeling (NTM) (Cao et al., 2015) uses neural networks to uncover latent topics within document collections, offering a more flexible and potentially more powerful alternative to classical probabilistic methods.

Variational Autoencoders (VAEs) (Srivastava and Sutton, 2017) combine the strengths of deep learning with probabilistic inference to discover latent topics in document collections and can be used as a component for NTM. The VAE architecture for topic modeling typically consists of an encoder network, a latent space, and a decoder network. The encoder maps input documents to a probability distribution in the latent space, usually parameterized by mean and variance vectors. This latent space represents the topic space, where each dimension corresponds to a topic. The decoder then takes samples from this latent distribution and attempts to reconstruct the original document. The training process of VAEs for topic modeling involves optimizing two objectives simultaneously: minimizing reconstruction error (how well the model can reconstruct the original document from its latent representation) and regularizing the latent space to approximate a prior distribution, typically a standard normal distribution. This regularization, achieved through the Kullback-Leibler divergence term in the loss function, encourages the model to learn a smooth, continuous topic space. The key advantage of VAEs is their ability to learn complex, non-linear mappings between documents and their latent topic representations, potentially capturing more nuanced topic structures than linear models. Another main benefit of using VAEs for topic modeling is their flexibility and scalability. Challenges of using VAEs are the potential for learning less interpretable topics compared to traditional methods and the need for careful hyperparameter tuning.

Researchers have proposed various implementations and extensions of NTMs, each addressing specific aspects of topic modeling. For instance, Zhu et al. (2020) incorporates attention mechanisms to improve topic coherence, and Liu et al. (2014) explores hierarchical structures to capture topic relationships at different levels of granularity.

2.1.3 Evaluation of topic modeling

The evaluation of topic models presents unique challenges that have spurred significant research interest. Unlike supervised learning tasks, topic modeling lacks a clear ground truth, making objective assessment difficult (Zhao et al., 2023). The interpretability and usefulness of topics can be subjective, varying based on domain expertise and research goals (Hoyle et al., 2021). Moreover, topic quality encompasses multiple aspects, including coherence, distinctiveness, and coverage, which may not always align. These challenges underscore the importance of robust evaluation methods in ensuring the validity and reliability of topic modeling results. There are two main approaches to evaluating topic models: human judgment and quantitative metrics.

Human judgment approaches involve the manual inspection and interpretation of topics by domain experts. The key advantage of these methods is their accuracy; the depth of insight provided by domain experts can lead to nuanced and comprehensive evaluations of the topics being studied. However, since these methods require human involvement, they take longer to complete and are not scalable.

A highly recommended method is topic intrusion (Chang et al., 2009). First, a topic model is trained on a corpus of documents to produce a set of topics. A sample of documents from the corpus is selected for evaluation. For each selected document, the top 3-4 topics most strongly associated with that document by the model are identified. An additional "intruder" topic that is not strongly associated with the document is randomly selected from the remaining topics. Human subjects see the document (usually just the title and a snippet) along with the set of 4-5 topics (the 3-4 real topics plus the intruder). The subjects are asked to identify which topic is the "intruder" that is not associated with the document. The success rate at which humans correctly identify the intruder topic is measured. Higher success rates indicate that the model's topic assignments align well with human intuition, suggesting the topics are coherent and meaningful.

Another approach is quantitative metrics, automated measures that can be computed at scale and can be done immediately. However, quantitative metrics are often not aligned with human judgment. Common metrics are perplexity and topic coherence scores.

Perplexity, or held-out likelihood, is a traditional metric (Azzopardi et al., 2003). This method involves splitting the dataset into a training set and a test set, called held-out documents. The topic model is trained and then used to predict the probability of the held-out documents in the test set. The likelihood of these held-out documents is calculated, usually as a logarithm, hence the term "held-out log-likelihood." Perplexity itself is computed as the exponential of the negative average log-likelihood per word, with lower values indicating better predictive performance. While this metric assesses how well the model generalizes to unseen data, it has notable limitations. Research has shown that perplexity does not necessarily correlate with human interpretability of topics, and in some cases, a negative correlation has been observed between perplexity and human judgment of topic quality (Chang et al., 2009). Furthermore, variations in sampling or approximation techniques across different topic models can complicate direct comparisons. Due to these constraints, while perplexity remains in use, it is often supplemented with other evaluation methods, particularly those involving human judgment or semantic coherence measures, to provide a more comprehensive assessment of topic model quality.

Topic coherence measures aim to quantify how semantically coherent or meaningful

the top words in a topic are (Rosner et al., 2014; Newman et al., 2010). Each topic is represented by its top N most probable words. For each topic, word pairs are formed from these top words. The semantic similarity or co-occurrence of these word pairs is then assessed. A confirmation measure is calculated for each word pair, usually based on word co-occurrence statistics from a reference corpus. The individual word pair scores are aggregated to produce an overall coherence score for the topic. Topic scores are typically averaged to give a coherence score for the entire topic model. Popular coherence measures include cosine similarity, normalized pointwise mutual information (NPMI), or document co-occurrence statistics. Higher coherence scores generally indicate more semantically coherent and human-interpretable topics. These measures tend to correlate well with human judgments of topic quality, making them valuable automated evaluation metrics for topic models.

2.1.4 Summary and choice

In summary, topic modeling has evolved from foundational probabilistic approaches like PLSA and LDA to more advanced methods incorporating neural networks and word embeddings. Probabilistic methods offer strong theoretical grounding but may require assumptions about topic independence or the number of topics. Non-probabilistic techniques like LSI and NMF provide interpretable results using linear algebra but can struggle with scalability or nuanced semantics. More recent neural and embedding-based models, such as those employing VAEs or contextualized embeddings like BERT, enable richer representations and improved topic quality, especially for complex or modern text corpora.

Evaluating topic models remains a complex task, requiring a balance between quantitative measures and human-centered assessments. While perplexity is widely used, it does not always align with human interpretability. Topic coherence metrics offer a stronger correlation with human judgments, and topic intrusion tests allow direct evaluation of how well topics align with human intuition.

Based on this review, I chose to use BERTopic for my topic modeling approach. BERTopic leverages transformer-based word embeddings, which offer rich contextual understanding, and combines them with clustering techniques to generate interpretable topics. For evaluating the generated topics, I employed both topic coherence and topic intrusion methods to ensure a balance between automated evaluation and human-centered interpretability. This combination offers a robust, state-of-the-art framework for uncovering and validating latent themes in text.

2.2 Background of Sentiment Analysis

Sentiment analysis involves determining the emotional tone of a piece of text by classifying the sentiment as positive, negative, or neutral. This is known as sentiment polarity. Through analyzing the words and phrases used in the text, sentiment analysis can uncover the underlying attitudes, opinions, and emotions of the writer. Additionally, this method can be employed to monitor changes in emotional expression over time.

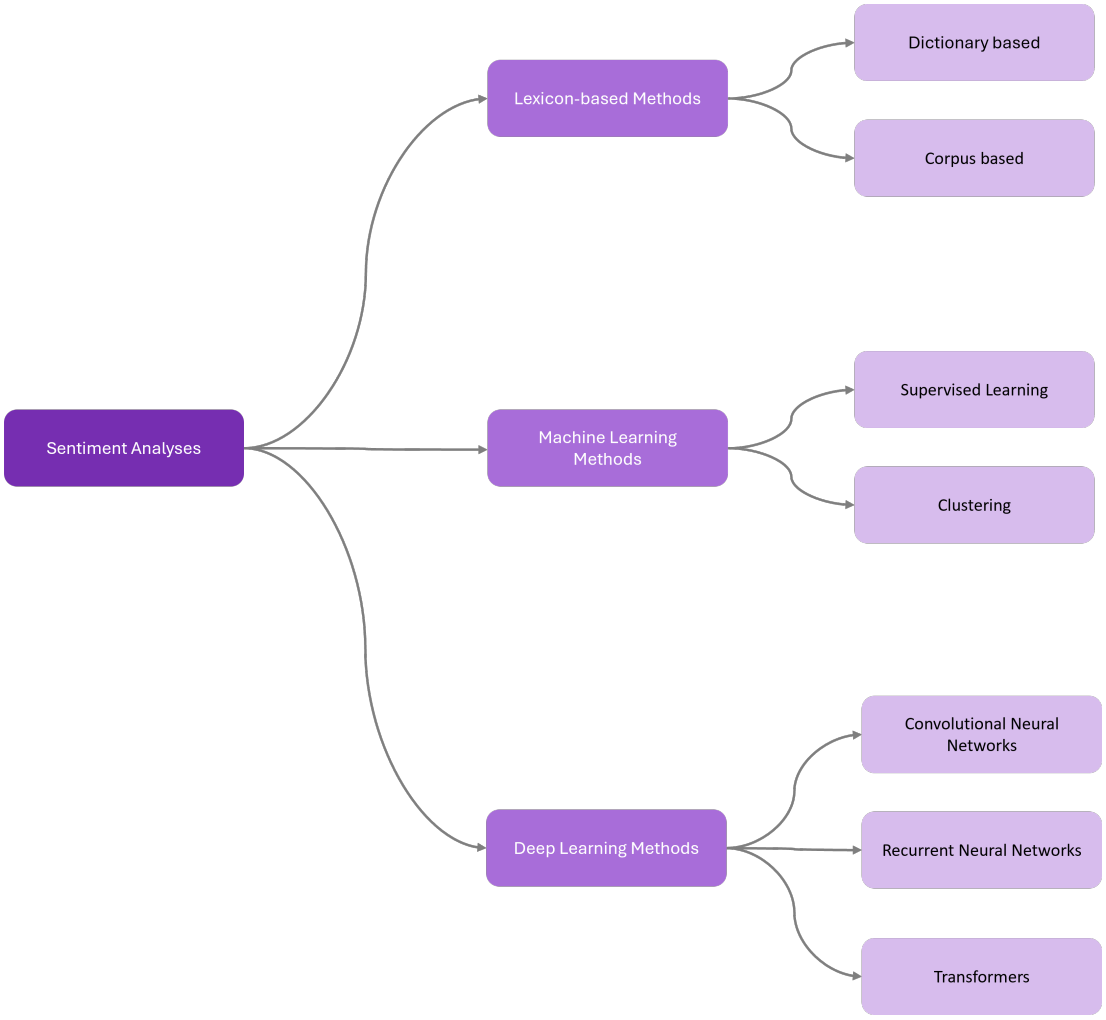


Figure 2.2: Hierarchy of Sentiment Analysis approaches.

2.2.1 Sentiment Analysis Methods

Figure 2.2 presents a hierarchy of different approaches for this task based on several surveys (Behdenna et al., 2024; Yadav and Vishwakarma, 2023; Nugraha et al., 2022; Liu, 2017). Each approach is further explained below.

Lexicon-based methods

Lexicon-based methods rely on predefined sentiment lexicons or dictionaries to determine the sentiment of the text.

An example of a dictionary-based method is SentiWordNet (Sebastiani and Esuli, 2006). The algorithm uses sentiment scores assigned to each synset (set of synonymous words) in the WordNet database. Specifically, SentiWordNet associates three numerical scores with each synset: positivity, negativity, and objectivity. Objectivity is calculated as 1 - (positivity + negativity). These scores indicate how positive, negative, or objective/neutral the terms within a synset are. The scores are automatically generated using semi-supervised machine learning algorithms and natural language processing techniques. For each word, the sentiment can be looked up and used for further

processing to determine the sentiment of the sentence, text, or document.

A corpus-based example is SentiStrength (Thelwall et al., 2010), a novel algorithm designed for sentiment strength detection in short, informal texts prevalent on social media platforms. SentiStrength's core is a lexicon-based system, augmented with a set of rules to handle informal language features such as emoticons, emphatic spelling, and negations. The algorithm's effectiveness is further enhanced through machine learning optimization of sentiment word strengths and rules. This approach employs a dual-scale system, separately measuring positive and negative sentiment on a 1-5 scale, offering a more nuanced analysis than traditional single-scale methods.

Machine Learning Approaches

Supervised learning approaches have been widely adopted in sentiment analysis, offering robust and effective methods for classifying text sentiment (Singh and Jaiswal, 2023). This approach relies on labeled training data to teach algorithms to recognize patterns associated with different sentiment polarities. Three examples of effective models are Support Vector Machines (SVM) (Pang and Lee, 2004), Naive Bayes classifier (Liu, 2022), and Decision trees (Singh and Tripathi, 2021). For most sentiment analysis methods, the choice of features is crucial to the model's effectiveness. Features can include a variety of textual elements that reflect different aspects of sentiment expression (Wankhade et al., 2022). These encompass lexical features such as n-grams (unigrams, bigrams, and trigrams) and pragmatic features that focus on the context of word usage. Emojis and emoticons are also important as they convey emotions, while punctuation marks can emphasize the strength of sentiment. Additionally, slang terms often express strong or nuanced underlying sentiments in a casual or conversational tone. Lastly, negations are vital as they can reverse sentiment polarity, and managing neutral sentiment and ambivalence adds depth to the analysis. Common approaches for the extraction for these features include term frequency calculations (such as TF-IDF) to measure word importance, Parts of Speech (POS) tagging to identify sentiment-bearing words (particularly adjectives), and the creation of Bag of Words (BoW) representations to capture overall word occurrence patterns (Wankhade et al., 2022).

However, labeled data may not always be available, and therefore, unsupervised learning approaches are also studied. For example, in the study of Hu et al. (2013), a K-means clustering algorithm is used on emotional signals in texts.

Deep Learning Approaches

Deep learning approaches have revolutionized sentiment analysis by leveraging neural networks to learn complex sentiment representations directly from data. Convolutional Neural Networks (CNNs) excel at capturing local features in text through word embeddings, proving effective for sentence-level sentiment classification (Ouyang et al., 2015). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) variants, have shown remarkable ability in modeling sequential dependencies in text (Tai et al., 2015; Zhou et al., 2016). The advent of transformer-based models has set new benchmarks in sentiment analysis tasks. BERT's bidirectional pre-training and fine-tuning approach enables it to capture context-dependent sentiment with unprecedented accuracy (Biswas et al., 2020). These deep learning methods have significantly

outperformed traditional approaches, offering more nuanced and accurate sentiment analysis across various domains and text types.

2.2.2 Evaluation of sentiment analysis

Human annotations involve having people manually label text samples with sentiment categories. This provides a "gold standard" dataset for evaluating automated sentiment analysis systems. To assess the reliability of human annotations, researchers measure inter-annotator agreement using metrics like Cohen's Kappa, which measures agreement between two annotators, Fleiss' Kappa, which measures agreement among three or more annotators, or Krippendorff's Alpha, which handles various types of data and missing values. With the "gold standard", metrics such as accuracy or mean absolute error can be calculated to quantify the effectiveness of the automated sentiment analysis system in replicating human judgments.

However, annotating a whole dataset is very extensive work. Therefore, other proxy metrics have been proposed over the years. Pang and Gimpel (2018) examine the distribution of sentiment scores. They compare sentiment score distributions to expected patterns, analyze the balance of positive, negative, and neutral sentiments, and look for anomalies or unexpected skews in the distribution. Xie et al. (2020) measure the consistency of sentiment predictions by applying data augmentation techniques (e.g., synonyms, paraphrasing) and comparing the sentiment scores before and after augmentation. Higher consistency indicates better performance. Others proposed using known models. For example, comparing sentiment predictions with established sentiment lexicons (Baccianella et al., 2010), and calculating agreement between system outputs and lexicon-based sentiment. Higher agreement suggests better alignment with human-curated resources. Another way is to use pre-trained language models to score the plausibility of sentiment predictions by calculating perplexity scores for sentiment-augmented texts (Radford et al., 2019). Lower perplexity indicates more natural and plausible sentiment assignments.

The methods mentioned use proxy metrics and may not accurately reflect true performance. One way to obtain more reliable results without annotating the entire dataset is to conduct a survey (Barnes et al., 2021; Mohammad, 2016). This involves having domain experts evaluate a sample of sentiment predictions, instead of the whole dataset for the gold truth. Unlike human annotations, which rely on labeling the entire dataset to create a "gold standard," this approach uses a smaller sample of data evaluated by experts to assess the performance of sentiment analysis systems. This method can be more resource-efficient while still providing valuable insights into the system's accuracy. There are a few different options in designing the survey, such as selecting a representative sample or ensuring inter-expert consistency, to ensure the reliability of the results.

2.2.3 Summary and choice

Sentiment analysis, the computational task of identifying the emotional tone of a text, typically in terms of positive, negative, or neutral sentiment, can be done with different techniques, such as lexicon-based, machine learning-based, and deep learning-based methods. Lexicon-based techniques use predefined dictionaries such as SentiWordNet and SentiStrength to assign sentiment scores to words or phrases. Machine learning approaches, including SVMs and Naive Bayes classifiers, rely on labeled data and engi-

neered features like n-grams, emojis, and negations. Unsupervised models like K-means are used for cases where labeled data is unavailable. Deep learning models, including CNNs, LSTMs, and BERT, are noted for their ability to learn complex sentiment patterns and outperform traditional methods.

Evaluating sentiment analysis models is a complicated task, since sentiment is subjective and human annotations are needed for secure performance metrics.

Based on this review, I decided to use TextBlob for my sentiment analysis. TextBlob employs a lexicon-based approach, making it quick and eliminating the need for labeled data. To evaluate the sentiment outputs, I chose to conduct a survey, which strikes a balance between human annotations and efficiency.

2.3 Using Historical texts

2.3.1 Challenges

When applying NLP to historical texts, several common challenges need to be addressed. Firstly, historical texts often have data quality issues due to errors introduced during the digitization process, such as inaccuracies from Optical Character Recognition techniques (Poncelas et al., 2020). Additionally, variations in spelling and changes in language over time make it challenging to consistently process historical texts (Marjanen et al., 2020). One significant challenge in analyzing historical texts is the lack of context, especially in short passages or fragments that may not provide sufficient information for accurate NLP methods. This issue is commonly encountered in NLP, but it is particularly critical for historical texts due to the way they may reflect earlier cultural norms or biases (Liu, 2024). Historical documents often contain language, idioms, or references that can be unfamiliar or have changed meaning over time. As a result, models might misunderstand or misinterpret the intended message without the necessary context. The social and cultural backdrop of historical texts can differ greatly from modern views. Therefore, the models need to be domain-specific. However, if not enough context is available, the models may not be accurate. Additionally, as research methods and our understanding of history evolve, interpretations of historical texts may change. Consequently, domain expertise is crucial for determining whether the NLP outcomes are meaningful or trivial. Lastly, historical corpora may exhibit uneven coverage across different time periods, leading to issues with data sparsity.

2.3.2 Related work

Multiple topic modeling techniques have been used for historical texts. Newman and Block (2006) applied LDA to 18th-century colonial American newspapers to uncover thematic trends. They identified topics related to commerce, politics, and social life, revealing patterns in colonial discourse. Mimno et al. (2009) used LDA on 19th-century novels to explore thematic differences across nationalities, genders, and periods. Their work demonstrated how topic modeling could reveal macro-level trends in literary history. They applied this model to English, French, and German novels, uncovering cross-linguistic thematic patterns. Nelson (2020) integrated topic modeling with network analysis to study the conceptual structure of British imperial thought in the 18th and 19th centuries. This combination allowed for a more nuanced understanding of how ideas were connected and disseminated. Klein and Eisenstein (2013) used topic modeling to analyze the writings of Thomas Jefferson, revealing how his interests and

concerns evolved throughout his political career, and providing a framework for analyzing cultural archives. The findings from this study indicate that scholars should be offered comprehensive support for visualizing and exploring topic model output. This support should be integrated with traditional workflows focused on gathering and improving relevant documents.

Several studies have applied sentiment analysis to historical texts, addressing unique challenges and developing specialized approaches. Krušić (2024) developed a sentiment-annotated corpus of 19th-century Austrian German newspapers, providing valuable resources for sentiment analysis of Austrian historical texts. In the realm of literary analysis, Allaith et al. (2023) created an annotated sentiment dataset for historical Danish and Norwegian texts, finding that pre-trained multilingual language models outperformed models trained on modern Danish and lexicon-based classifiers. Sprugnoli et al. (2016) focused on political writings, analyzing the works of Alcide De Gasperi and evaluating existing lexical resources for sentiment analysis on historical texts. They emphasized the importance of domain adaptation and explored crowdsourced annotation for obtaining contextual polarity.

These studies collectively highlight the challenges and potential of NLP for historical texts, emphasizing the need for specialized approaches and resources to address the unique characteristics of historical language and context.

Chapter 3

Methodology

This chapter outlines the methodological frameworks employed in this study to explore the feasibility and usefulness of computational techniques for analyzing the popularity of art trends in historical texts. Rather than aiming to compare multiple tools, the study takes an exploratory approach, applying selected methods to determine whether they can meaningfully support the historical analysis of early modern English sources.

At each stage, a single model or technique was selected to carry out the analysis. This decision reflects the exploratory nature of the study, which aims to assess the feasibility and practical value of applying computational methods to historical texts. While no comparative evaluation of multiple tools is undertaken, the selected models were chosen based on their proven performance or suitability for the linguistic and conceptual challenges posed by early modern English sources.

Additionally, recognizing that topic modeling may not always capture specific, predefined art trends, an alternative method based on keyword expansion and semantic text matching is introduced. This approach uses word embeddings and accommodates historical spelling variation to improve coverage of relevant references and terminology.

By combining computational techniques with domain expertise, this methodology provides a structured approach to uncovering insights from historical texts. The following sections elaborate on each step of the framework, detailing the processes and justifications behind the methodological choices.

3.1 Data

The texts used for this study are obtained using the resources Early English Books Online (EEBO) from ProQuest and Eighteenth Century Collections Online (ECCO) from Gale Cengage. The selection of these resources is based on artistic movements that occurred between 1500 and 1800; one of these, the landscape genre, will be examined to illustrate the framework. Specifically, EEBO includes page images of almost every work printed in the British Isles, North America, and English works printed elsewhere from 1470-1700, and has received contributions from over 200 libraries worldwide (ProQuest). ECCO includes significant English-language and foreign-language titles printed in the United Kingdom during the 18th century, along with thousands of important works from the Americas (Gale Cengage).

The Text Creation Partnership generated fully searchable, SGML/XML-encoded texts corresponding to books from the EEBO and ECCO (Text Creation Partnership, a,b) by using Optical Character Recognition techniques for ECCO and manually tran-

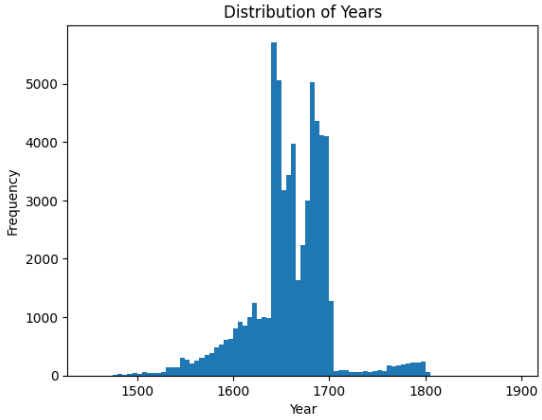


Figure 3.1: Frequency of texts in EEBO and ECCO over the years.

scribing for EEBO. EEBO-TCP and ECCO-TCP, while unstructured, are valuable resources for Early Modern English scholars. They provide a vast collection of texts that allow for a comprehensive view of the language over time. 301 documents were excluded from this study because they did not have a date that could be extracted. In total, 62,498 documents were used, the distribution of documents over the years is visible in Figure 3.1.

It is important to note that the OCR process is not entirely free of inaccuracies. Consequently, certain words within the digitized documents may not accurately reflect their intended form. For instance, the letters 'i' and 'l' are often subject to confusion. Given that this research primarily focuses on methodological analysis rather than textual accuracy at the individual word level, minor OCR discrepancies are accepted as a tolerable limitation. While systematic errors could theoretically affect certain patterns, the overall trends observed in the study are unlikely to be significantly distorted by such variations. A targeted evaluation of OCR accuracy, outside of the general error analysis, would be ideal, but it falls outside the scope of this research. This study acknowledges the potential presence of OCR-related noise.

3.1.1 Data selection and preproccessing

The purpose of this study is to identify trends in art-related documents. Several steps will be taken to extract texts specifically related to art. Some of these steps may result in the removal of certain art-related texts. I believe that prioritizing a clean set of paragraphs is more valuable than striving for completeness. While individual paragraphs may contain errors due to optical character recognition (OCR), this approach ensures that the overall collection remains focused and relevant, as all paragraphs pertain to the theme of art. Furthermore, I have chosen to look at paragraphs. Analyzing full documents can be too coarse-grained because they often contain multiple themes and ideas, making it hard to identify specific arguments or nuances. Additionally, a full document can have sections about art and others that do not, which would corrupt the dataset. In contrast, examining a single sentence can be too fine-grained, as it may lack the necessary context to understand its implications fully, such as topics or sentiments. Paragraphs offer a balanced approach, typically focusing on a single idea while providing enough context to appreciate the surrounding details, making them ideal for topic and sentiment analysis for this purpose.

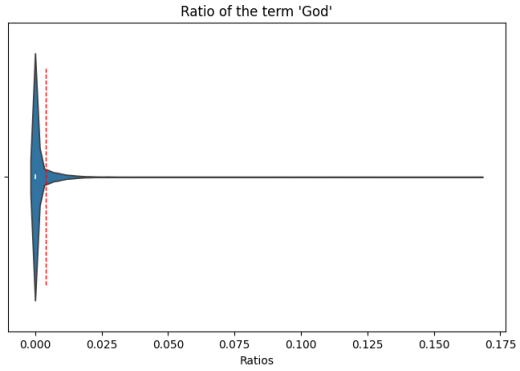


Figure 3.2: Ratio of the term 'God' in the paragraphs. The red dotted line indicates the cut made for determining which texts are too religious.

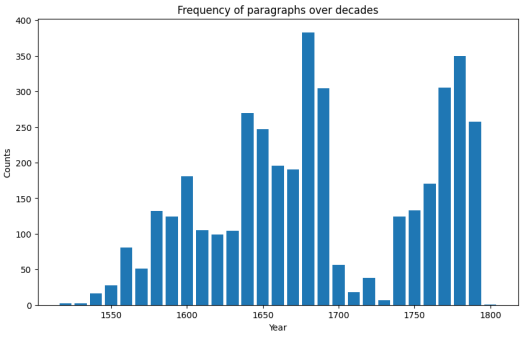


Figure 3.3: Frequency of the extracted paragraphs over decades.

- I excluded scripts for plays from the corpus because I can only extract spoken sentences from them, which makes them irrelevant to this analysis. Consequently, 1,336 documents were excluded during this step.

- Since the corpus is historical, it contains a significant number of religious writings, which can introduce biases. To mitigate this issue, I decided to remove the most religious documents by establishing a threshold for the word "God" within the sections of the documents, as shown in Figure 3.2. I set the ratio at 0.0039 for any section of the document, as this clearly defines the distribution. As a result of this step, 1,763 documents were excluded.

- The term "art" refers not only to artistic creations but also to exceptional skills and craftsmanship. As a result, I cannot rely on the word "art" to filter relevant content. Instead, I utilized a list of 8 other keywords that signify a text related to art. I used these keywords to extract paragraphs from the documents that included at least one of these terms. In total, 3973 paragraphs concerning art were extracted.

Each paragraph is tokenized using spaCy. I utilized a standard English model because I was unable to find a model specifically designed for early modern English. Although I did find models for Old English and Middle English (Johnson et al., 2014–2021), early modern English resembles the current English more closely than those varieties do. However, this may result in less tokenization than expected, since some words may be unknown to the model. After inspection, the outcome from spaCy appeared reasonable and better than the Old English version.

Figure 3.3 shows the distribution of the extracted paragraphs over the decades. Between 1700 and 1730, there is a notably low number of paragraphs, which may affect the analysis for these decades. Additionally, the year 1800 has only one paragraph, making it unreliable.

3.2 Framework for Popularity of Art Trends

To assess art trends and their popularity, I propose a framework that integrates topic modeling with sentiment analysis to facilitate trend analysis. This section will explain these steps, including the choices and methods employed.

3.2.1 Topic Modeling

For topic modeling, a model called BERTopic is used, which leverages BERT embeddings and clustering techniques to identify coherent topics within the corpus (Grootendorst, 2022). They reported state-of-the-art results, so I expect this technique to perform well on this corpus as well. The process can be summarized in the following steps. First, each document d_i is transformed into a dense vector representation $\mathbf{e}_i$ using a pre-trained BERT model. In this study, MacBERTh was compared to standard BERT to identify more relevant topics (Manjavacas Arévalo and Fonteyn, 2021). MacBERTh is trained on documents from William Shakespeare (1564 - 1616), written in Early Modern English, used in a period that spans from 1500 to 1700. Since the paragraphs are dated from 1550 to 1800, they are more likely to align with the MacBERTh style compared to standard BERT models. Next, the high-dimensional embeddings are reduced using UMAP (Uniform Manifold Approximation and Projection). The reduced embeddings are then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise. Lastly, for each cluster, the most representative terms are extracted using a KeyBERT-inspired method, which identifies semantically meaningful keywords by comparing candidate word embeddings with the average embedding of documents in the topic cluster. This differs from the default c-TF-IDF representation by prioritizing semantic relevance over frequency.

In addition, to improve interpretability and reduce redundancy among similar topics, an automatic topic reduction step is applied. This step uses the same KeyBERT-inspired embeddings of topic keywords to calculate semantic similarity between topics and iteratively merges the most similar ones until a suitable level of distinctiveness is reached.

To evaluate the quality of the generated topics, topic coherence, topic intrusion, and topic alignment are used. Topic Coherence is measured by the normalized pointwise mutual information (NPMI) between the top N words of a topic:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

where $P(w_i, w_j)$ is the probability of words w_i and w_j co-occurring, and $P(w_i)$ is the probability of word w_i occurring.

Topic intrusion is assessed through a survey. In this survey, human evaluators are given a set of N words, where $N-1$ words belong to one topic and 1 word (the intruder) comes from a different topic. The following formula measures the evaluators' ability to identify the intruder:

$$\text{Intruder Detection} = \frac{\text{Number of correct identifications}}{\text{Total number of evaluations}}$$

Topic alignment is also assessed through a survey. Human annotators are presented with paragraphs and asked to identify the topic to which each paragraph belongs. The

participants have five different topics to choose from. To increase the difficulty of the task and more accurately evaluate the alignment of the topics, the options are not completely random. If a paragraph contains a word that is associated with another topic, meaning the word is present in that topic as well, that topic will be included as one of the options. This approach is designed to ensure that the topics are relevant.

More about the set-up of the surveys is available in Chapter 4.

3.2.2 Sentiment Analysis

For the sentiment analysis, I chose to use an off-the-shelf tool, as training a new model would require a gold-standard dataset. Manually annotating a sufficiently large dataset for accurate training would demand significant effort from domain experts. Therefore, I opted for TextBlob (Loria, 2014), a simplified natural language processing library that provides an accessible and straightforward sentiment analysis tool. Unlike deep learning-based sentiment models, which often require fine-tuning on domain-specific data, TextBlob uses a lexicon-based approach, making it more suitable for general sentiment classification without additional training. While it does not specifically account for historical or art-related texts, its rule-based method ensures a consistent sentiment scoring mechanism that does not rely on contemporary language trends. Due to time constraints, exploring domain adaptation on more advanced models was not feasible for this stage of the framework.

TextBlob assesses the sentiment of textual data by analyzing the polarity and subjectivity. The polarity score ranges from -1 to 1, representing the intensity of negative (-1) to neutral (0) to positive (1) sentiment. Subjectivity measures the text's objectivity or subjectivity, ranging from 0 to 1 respectively. I used the polarity scores for the trend analysis.

To evaluate how well the sentiment analysis works on this data, a survey is taken. The participants are presented with art-related paragraphs and are asked to evaluate the sentiment from -3 to 3. The Root Mean Square Error will be computed on the averaged normalized survey findings and a Spearson correlation test will be conducted to evaluate the model's accuracy.

3.2.3 Trend Analysis

The trend analysis component of the framework involves tracking the frequency and sentiment of topics over time. Specifically, I will plot the frequency of topics over time to identify patterns and study emerging, declining, and cyclical trends. To assess sentiment shifts of these trends, I will integrate the sentiment analysis component, examining how the sentiment associated with specific art trends changes over time.

This allows for the identification of:

1. Emerging trends: Topics that show a sudden or gradual increase in frequency
2. Declining trends: Topics that demonstrate a decrease in mentions over time
3. Cyclical trends: Topics that show periodic patterns of popularity
4. Sentiment shifts: Changes in the overall sentiment towards specific art trends, as analyzed in conjunction with sentiment analysis

3.3 Alternative Framework for Specific Trend Analysis

Since the framework relies on automatically generated topics, some trends may not be included in these topics. When art historians are aware of a specific trend that should be addressed but is not captured during this step, an alternative approach becomes necessary. Therefore, a different framework is proposed. Instead of using topic modeling, a keyword search will be employed.

The process begins with a keyword or latent topic provided by an art historian. Since not all paragraphs that are related to this keyword or topic has those words in them, a more advanced search is neccesary. To account for variations in language and to ensure comprehensive coverage, query expansion using word embeddings is a common method (Diaz et al., 2016; Kuzi et al., 2016). For the first try a new Word2Vec model has been trained on the TinyShakespeare dataset (Karpathy, 2015) to generate context-appropriate word embeddings. Word2Vec maps words to a high-dimensional vector space where semantically similar words are positioned closer together (Mikolov et al., 2013). The dataset was so sparse in terms of nature and landscape that the results were not interpretable. Therefore, I decided to switch to using the GloVe model (Pennington et al., 2014). The project offers several pretrained models, and I chose the one trained on Wikipedia 2014 and Gigaword 5, which utilizes 100-dimensional vectors. The other models are trained on Common Crawl (which primarily includes news articles and blogs) or Twitter, making my chosen model more relevant to the specific domain of art related paragraphs.

Given a word w, the top 3 related terms R are identified using cosine similarity:

$$R = \{r : \cos(\vec{w}, \vec{r}) \geq \theta\}[0:3],$$

where $\vec{w}$ and $\vec{r}$ are the vector representations of words, and θ is a similarity threshold, set on 0.5.

To account for spelling variations in Early Modern English, I consulted the Oxford English Dictionary to further expand the query with historical spellings (Oxford English Dictionary, 1857). This is a manual step in the process, which seems manageable.

Text Matching

Each paragraph p in the corpus and each word q in the expanded query set Q are checked for matches. To mitigate OCR errors, the Levenshtein distance $L(q, w)$ is employed between query words and words in the paragraph:

$$M_p = \{q \in Q : \exists w \in p, L(q, w) \leq \delta\},$$

where $L(q, w)$ is the Levenshtein distance and δ is a threshold for an acceptable string distance, set at 0.75. The smallest number of single-character modifications needed to change one string into another is known as the Levenshtein distance between two strings. The allowed edits are insertion of a character, deletion of a character and substitution of a character. For OCR correction and accuracy enhancements, this distance is frequently used (Vukatana, 2022; Myka and Güntzer, 1996).

Chapter 4

Surveys

Some surveys needed to be conducted to assess the accuracy of the models presented in Chapter 5. Three surveys were conducted. The first survey was designed to evaluate the accuracy of the sentiment model, the second survey was to evaluate the coherence of the topics, using topic intrusion, and the last survey aimed to evaluate the assignment of topics to paragraphs. However, since this last survey received only three responses, it is not suitable for use as a test set. Nonetheless, in this chapter, I will explain the structure of the three surveys and evaluate the annotators. A necessary step before assessing the sentiment- and topic-model with the results of the surveys. Maintaining the reliability of annotations in a crowdsourced task is essential for ensuring the quality of the resulting metrics. To assess the accuracy of survey data, an inter-annotator agreement analysis is conducted for each survey.

4.1 Development set

To create the surveys, a development set of 60 paragraphs was created. It is essential that these paragraphs represent the dataset by including enough early modern English to be relevant, yet still remain clear and understandable for assessment, without being too lengthy for the survey. As a result, paragraphs were randomly selected from the dataset, specifically those that had a character length between 100 and 180. Each paragraph was carefully reviewed to ensure it was clear enough for presentation to others.[1] The full development set is shown in appendix B with the assigned topic and sentiment from the models.

4.2 Sentiment Survey

The first survey was designed to evaluate the efficacy of the sentiment model and was structured as follows:

1. Participants commenced by responding to two questions aimed at assessing their knowledge in the domains of text mining/Natural Language Processing (NLP) and art history. They were instructed to provide a self-assessment score ranging from 1 to 5, with 5 denoting a master level of expertise and 1 indicating a complete

[1] There has been no tracking of the texts that were discarded; however, no more than ten texts were removed. This method may introduce bias into the results, as unclear paragraphs are more likely to result in errors by humans and longer ones as well since they require more focus from the participants.

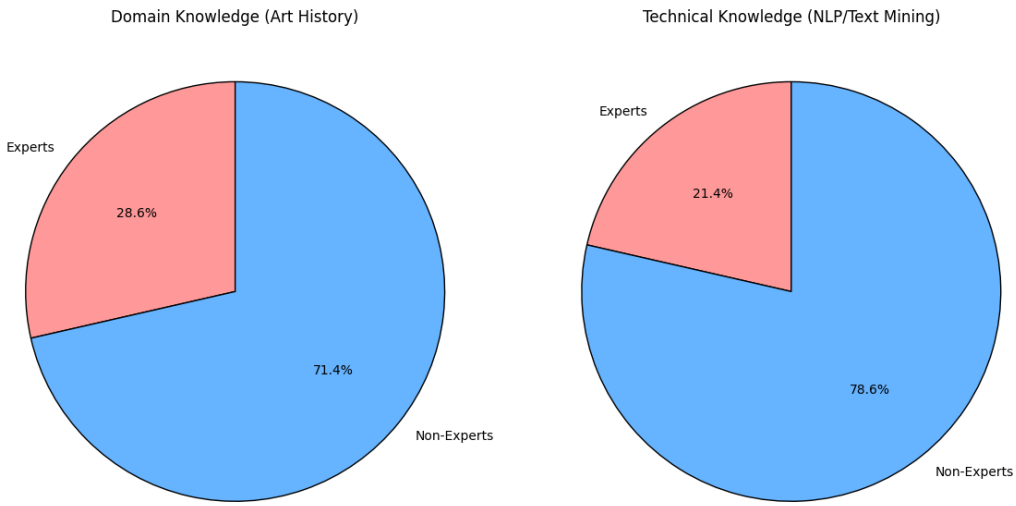


Figure 4.1: Distrubuation of domain and technical experts for the Sentiment Analysis Survey.

lack of knowledge. Individuals who rated themselves as 4 or 5 in the art history domain were classified as domain experts, while those who rated themselves a score of 4 or 5 in text mining/NLP were designated as technical experts.

2. All participants then rated the sentiment of a fixed control set of 10 paragraphs on a scale from -3 (very negative) to +3 (very positive), which allowed for inter-annotator agreement analysis.

3. After the control set, participants were randomly assigned to additional sets of 10 texts, with the option to complete up to 60 questions in total. After each set of 10, participants got the option to stop the survey. The randomized assignment ensured diverse responses while getting enough data for measurements.

The survey got 12 responses, the distribution of experts is shown in Figure 4.1.

4.2.1 Inter-Annotator Agreement

In the sentiment survey, I follow two steps to assess inter-annotator agreement. First, I identify unreliable annotators and remove them, leading to a higher *Krippendorff's Alpha*. Finally, I compare experts and non-experts regarding both domain knowledge and technical knowledge to determine if bias influences their responses.

Identifying Unreliable Annotators

To systematically identify and eliminate unreliable annotators in the context of sentiment analysis utilizing Likert-scale-based data, I employ *Krippendorff's Alpha* (α), which is a widely recognized metric for assessing inter-annotator agreement in ordinal data (Artstein and Poesio, 2008).

The procedure for identifying unreliable annotators encompasses the computation of Krippendorff's Alpha for the entire set of annotators, followed by an iterative process

of removing each annotator to examine the impact of their removal on the overall agreement score. The difference between those is also called the Raters Vitality, a quality metric for annotators (McCulloh et al., 2018). The steps in this analysis are as follows:

1. **Normalization:** Recognizing that annotators may utilize the scale differently—some employing the full spectrum while others tend to cluster towards the midpoint—a Z-normalization procedure is applied for each annotator.

2. **Initial Calculation of Krippendorff's Alpha:** The baseline Krippendorff's Alpha is computed using all annotators, serving as a reference point for subsequent analyses.

3. **Iterative Removal of Annotators:** Each annotator is systematically removed from the dataset one at a time. Following the exclusion of an annotator, Krippendorff's Alpha is recalculated for the remaining annotators.

4. **Identification of Unreliable Annotators:** If the Rater Vitality is less than -0.05, the annotator will be flagged as unreliable. McCulloh et al. (2018) suggests removing annotators with a negative Rater Vitality, but this approach may be excessive. Therefore, this threshold has been established to ensure that only those annotators whose removal would lead to a significant decrease in agreement are identified and eliminated.

This methodology facilitates the identification of annotators who may introduce inconsistency or noise into the dataset, potentially resulting in diminished overall agreement scores. By excising such annotators from the analysis, the remaining data is likely to exhibit greater coherence and, as a result, lead to more reliable evaluations. While Krippendorff's Alpha is an invaluable tool for measuring inter-annotator agreement, it is crucial to recognize that the exclusion of annotators based solely on agreement metrics may not always be warranted. Given that annotators may possess varying interpretations of the Likert scale, a certain degree of disagreement is to be anticipated in inherently subjective tasks such as sentiment analysis (Johnson and Creech, 1995; Gonzalez et al., 2020). Nevertheless, when a substantial improvement in agreement is observed post-removal, this suggests that the removed annotator's responses may not be consistent with the majority, thereby potentially skewing the findings. The calculations indicated that one annotator should be removed. After a manual check, their response appeared problematic, leading to the annotator's exclusion from the analysis.

Comparison of Expert and Non-Expert Annotators

Annotators may have two kinds of knowledge that could bias their responses. The first one is art historian knowledge, which would indicate a better understanding of the paragraphs and therefore a better annotation. The second one is text mining knowledge, which could bias the response since these annotators understand how a machine looks at the paragraphs and come up with an answer by that. To assess whether expert annotators, defined as those with a self-reported knowledge score of 4 or 5, differ significantly from non-expert annotators in their Likert scale responses, I employed the Two One-Sided Tests (TOST) procedure (Lakens, 2017). In TOST, one defines an equivalence margin Δ (here ± 2 points on the Likert scale) within which

	Art History Experts				Text Mining Experts			
Question	p_{low}	p_{high}	Difference	Equivalent	p_{low}	p_{high}	Difference	Equivalent
1	0.156	0.364	−0.556	Not	0.000	0.000	0.800	
2	0.056	0.465	−1.139	Not	0.173	0.133	0.167	Not
3	0.002	0.010	−0.528		0.085	0.013	0.933	
4	0.076	0.218	−0.528	Not	0.028	0.025	0.067	
5	0.000	0.000	−0.333		0.011	0.130	−1.167	Not
6	0.027	0.121	−0.611	Not	0.090	0.012	0.967	
7	0.170	0.256	−0.250	Not	0.048	0.017	0.533	
8	0.005	0.003	0.194		0.028	0.025	0.067	
9	0.017	0.002	0.806		0.108	0.224	−0.467	Not
10	0.021	0.003	0.722		0.000	0.000	0.000	

Table 4.1: TOST Results results comparing art historian experts and Text Mining/NLP experts with non-experts.

differences are deemed negligible. For each question, I conducted two Welch-adjusted one-sided t-tests:

$$H_{0L} : \mu_{\text{expert}} - \mu_{\text{non}} \leq -\Delta \quad \text{vs.} \quad H_{aL} : \mu_{\text{expert}} - \mu_{\text{non}} > -\Delta,$$

$$H_{0U} : \mu_{\text{expert}} - \mu_{\text{non}} \geq +\Delta \quad \text{vs.} \quad H_{aU} : \mu_{\text{expert}} - \mu_{\text{non}} < +\Delta.$$

If both tests reject at $\alpha = 0.05$, the mean responses are declared statistically equivalent within $\pm\Delta$. I carried out all analyses after excluding the unreliable annotator. Table 4.1 reports the one-sided p-values, observed mean differences, and equivalence conclusions for each question.

The results indicate that, for art-history experts, five out of ten questions met the TOST equivalence criteria (both one-sided p-values < 0.05). The other five items failed to reach equivalence, suggesting that art-historical expertise may systematically shift ratings on those questions. Because my pool of art-history annotators was small, I chose to retain them despite this partial non-equivalence in order to avoid a drastic sample-size reduction. However I do note that art historic knowledge has a bias.

By contrast, text mining experts met equivalence on seven out of ten questions—only three items fell outside the 2-point margin, demonstrating strong agreement with non-experts. Given this high concordance, I included text-mining experts without reservation in all subsequent analyses.

4.3 Topic Intrusion

The second survey aims to evaluate topic modeling effectiveness.

1. The participants began by responding to the two questions aimed at assessing their knowledge in the domains of text mining/Natural Language Processing (NLP) and art history. The same questions and logic were used as in the previous survey.

2. The participants were tasked with identifying the intruder word within a latent topic, where each topic comprised four keywords and one word that was selected from another topic. Participants were instructed to select the word they perceived as the intruder.

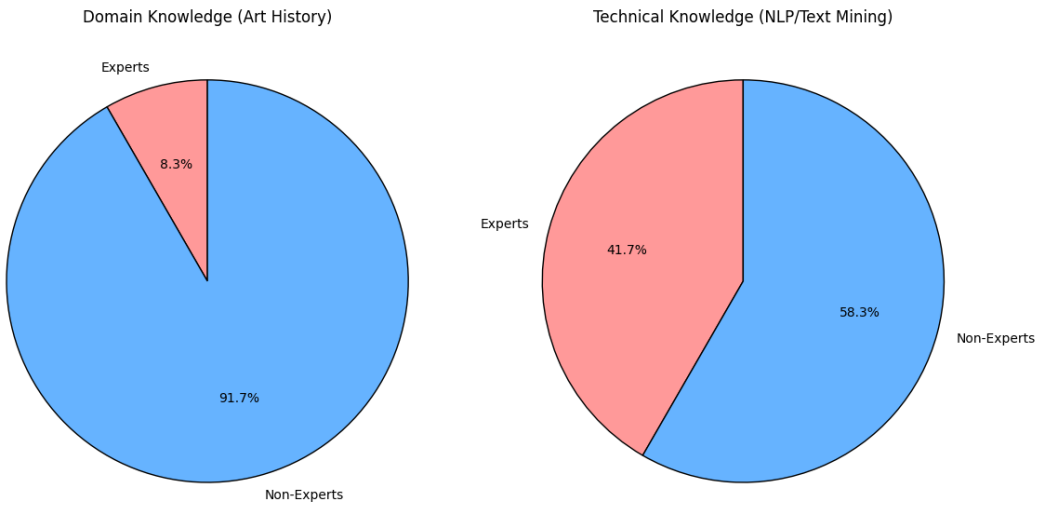


Figure 4.2: Distribuation of domain and technical experts for the Topic Intrusion Survey

. The survey got 12 responses, the distribution of experts is shown in Figure 4.2.

4.3.1 Inter-Annotator Agreement

For the topic intrusion survey, there are two steps involved in assessing inter-annotator agreement. First, I identify unreliable annotators and remove those whose assessments do not align with the majority vote. Next, I will compare the input from experts and non-experts in terms of both domain knowledge and technical knowledge to investigate whether bias has influenced their responses.

Identifying Unreliable Annotators

To assess the reliability of individual annotators, a consistency check was performed using majority voting as the reference standard. For each topic intrusion question, the majority vote was determined based on the most frequently selected answer among all annotators. Each annotator's responses were then compared to this majority vote, and their agreement rate was calculated. Annotators with consistently low agreement rates would have been considered unreliable. However, upon analysis using a threshold of 50%, no annotators exhibited a level of disagreement substantial enough to warrant removal. Thus, all responses were retained for further evaluation.

Comparison of Expert and Non-Expert Annotators

The Chi-square test examined the significant differences between the responses of domain experts and non-experts, and technical experts and non-experts, regarding topic intrusion questions.

For domain experts, the Chi-square statistic was 11.69, with a p-value of 0.98, and the degrees of freedom were 24. Since the p-value exceeds the significance threshold of 0.05, the null hypothesis is not rejected, suggesting no significant difference in responses to the topic intrusion questions.

For technical experts, the Chi-square statistic was 16.37, with a p-value of 0.87, and degrees of freedom were 24. The p-value (0.87) also exceeds the 0.05 threshold, leading to the non-rejection of the null hypothesis, which indicates no significant difference in responses.

Overall, there is no evidence of systematic bias between experts and non-experts in evaluating the topic intrusion task. Thus, it is unnecessary to exclude annotators based on their expertise.

4.4 Topic Alignment Survey

The last survey, the topic alignment survey, was designed to evaluate how well topics were assigned to paragraphs. While the topics may be coherent, this does not necessarily mean that their assignment is accurate. Only three responses were collected, which makes this test set unreliable. Unfortunately, due to time constraints, it was not possible to recruit more annotators. Furthermore, the annotators were not domain experts, making it difficult to determine if they lacked essential knowledge for accurate annotations. Nevertheless, I will report the structure and agreements of the survey. The structure was very similar to the sentiment survey:

1. Participants again started by answering the two questions about their expertise in NLP and Art History.

2. All participants then got a fixed control set of 10 paragraphs, the same set used in the sentiment analyse, and were asked to select one of the five options to assign the latent topic of the paragraph. One of the options corresponded to the topic identified by the model, while the remaining four options were also generated by the model but were not entirely random. If a paragraph included a word associated with another topic, that topic would be included as one of the options. If it was not possible to fill all options using this method, the remaining options were selected randomly.

3. After the control set, participants were randomly assigned to additional sets of 10 texts, with the option to complete up to 60 questions in total. After each set of 10, participants got the option to stop the survey. The randomized assignment ensured diverse responses throughout the remaining 50 questions.

4.4.1 Inner Annotator Agreement

Out of the 10 paragraphs in the control set, the annotators only reached an agreement on two of them. For six paragraphs, two annotators agreed, while the remaining two paragraphs received different topics from all annotators. After analyzing the removal of one of the annotators to see if that would improve agreement, the conclusion was that this approach did not enhance the level of consensus. In conclusion, there is insufficient data and the consensus is too weak to fully trust the results. To still gain a better understanding of the performance related to the topic alignment, a manual error analysis is conducted on the development set. The findings from this analysis is presented in the next chapter.

4.5 Conclusion

This chapter outlined the surveys conducted for the evaluation process for the sentiment analysis model and topic modeling. The first survey, focused on sentiment analysis, needed a filtering of one unreliable annotator. After the filtering, the tests showed that their is some bias for domain experts. However, since the test size would be too small by removing the non-experts, I decided to let it be. However, this bias is noted for any conclusions. With only three statistically significant difference across ten questions, the technical experts did not introduce a significant bias, allowing all remaining responses to be used in the evaluation of sentiment predictions. The second survey, evaluating topic coherence through a topic intrusion task, showed strong inter-annotator agreement. No annotators were removed, as their responses were consistent with majority voting. Additionally, statistical tests indicated no significant differences between expert and non-expert responses, confirming that expertise did not influence performance in this task. The final survey, assessing the alignment of topics to paragraphs, was ultimately inconclusive due to the limited number of responses and low agreement levels among annotators. Despite attempts to analyze the data, the lack of consensus rendered the results unreliable for further evaluation.

Chapter 5

Results

In this chapter, I will present the results from two frameworks. I will start with the first framework in Section 5.1, which aims to identify trends in art history. In Section 5.1.1, I will discuss the extracted topics. Both models will be evaluated in Section 5.1.3, where I will include coherence scores, the results of the topic intrusion survey, and an error analysis for the topic model. Additionally, I will analyze the sentiment analysis in conjunction with the survey results and an error analysis for the sentiment model. Next, in Section 5.1.2, I will explore the trends generated by the general framework, evaluating their relevance and logic within the context of established art history as a user with art historical knowledge. Finally, in Section 5.2, I will discuss the results of the specific framework focused on landscape paintings and analyze the differences between the two frameworks.

5.1 General Framework

The general framework aims to identify trends within the dataset. This section will present the resulting trends and evaluate the models employed in this framework.

5.1.1 Topics

With the BERTopic, the final model got 29 topics and one 'remaining' topic with miscellaneous paragraphs listed in Table 5.1. Different parameter settings were compared, including which autoencoder model to use, the approach, and whether or not to delete stopwords beforehand. The model with the highest cosine similarity score was selected. As stated before in the methodology, the BERTopic model was initialized using MacBERTh (Manjavacas Arévalo and Fonteyn, 2021) as the embedding model, which

-1: unto, therefore, selfe, nature	1: poetry, words, poet, manner	2: worship, doctrine, bishops, parliament
3: romans, roman, latin, greek	4: upon, speak, wise, onely	5: wordes, euerye, feare, thinke
6: royal, rafael, eminent, seville	7: parliament, lords, scots, commons	8: cathedral, church, statues, temple
9: ships, sail, ship, boats	10: king, kings, princes, noble	11: finely, flower, delicately, piece
12: dulnesse, howsoeuer, euery, shortnesse	13: prince, princes, mistress, king	14: raphaels, raphael, manuscript, artists
15: loue, sighte, blinde, looke	16: kingdom, royall, persian, townes	17: appearance, faces, complexions, natives
18: houses, decoration, ornaments, walls	19: heretic, punish, tongues, offence	20: bishop, royal, noble, archbishop
21: lande, sayled, captaine, foote	22: commodities, goods, silks, importation	23: landskip, finely, copper, dutch
24: authority, laws, law, warrant	25: seene, seate, capsterne, foure	26: ought, virtue, authority, judiciary
27: women, woman, filthines, herculean	28: vision, eyes, retina, sight	29: misery, desire, wrought, unto

Table 5.1: Latent topics created with BERTopic.

provides contextualized sentence embeddings that reflect the same language usage as the paragraphs, enhancing topic clustering. Before applying the model, stopwords from nltk were removed, along with a custom list of terms, including archaic pronouns (thee, th, thy, thou, shall) and the words related to art used for selecting the paragraphs (art, arts, painted, painters, painting, etc.), ensuring that topic extraction was not biased by frequently occurring but non-informative words. The KeyBERT-inspired representation model was chosen to extract the most representative words for each topic by prioritizing semantically meaningful terms rather than just frequently occurring ones. After fitting and transforming the documents, automatic topic reduction was applied to merge similar topics and refine the overall topic structure. This approach helps reduce redundancy, ensuring that the final set of topics is both distinct and interpretable.

The cosine similarity coherence scores were calculated using GenSim. This metric evaluates the semantic similarity between words in a topic based on a sliding window approach, called normalized pointwise mutual information (NPMI), and cosine similarity. Higher cosine similarity coherence scores indicate better topic coherence, meaning the words within each topic are more semantically related. Typically, coherence scores range from 0 to 1, with scores above 0.5 indicating well-formed topics, while lower scores suggest a lack of clear thematic structure.

The resulting coherence score for the generated topics was 0.492, which suggests a moderate level of topic coherence. While this score indicates that the topics capture meaningful structure within the dataset, there is room for improvement. A higher coherence score would indicate a better semantic similarity among words within each topic.

What is interesting to see is that several topics share similar themes. For instance, *10: king, kings, princes, noble*, *13: prince, princes, mistress, king*, and *16: kingdom, royall, persian, townes* revolve around monarchy, making them closely related. Similarly, *7: parliament, lords, scots, commons* and *20: bishop, royal, noble, archbishop* explore aspects of authority, though the former focuses on legal bodies and the latter on noble figures. Another notable pair is *6: royal, rafael, eminent, seville* and *14: raphaels, raphael, manuscript, artists*, which share a connection through Renaissance art, particularly highlighting the artist Raphael. These overlapping topics suggest shared contexts and are therefore not easy to distinguish from each other.

Other several individual topics stand out due to their clear and specific themes. For example, *9: Ships, sail, ship, boats* highlight maritime activities, and *8: cathedral, church, statues, temple* is focused on religious buildings. Similarly, *22: commodities, goods, silks, importation* reflects trade and economic exchange, and *23: landskip, finely, copper, dutch* emphasizes Dutch landscapes and craftsmanship. The topic *18: houses, decoration, ornaments, walls* is interpretable as well, focusing on interior design. Topic *29: misery, desire, wrought, unto* brings together themes of emotion and internal conflict, presenting a topic that is interpretable despite its abstract language.

Most of the topics are less interpretable due to their lack of a coherent theme or meaningful connection between the words. For instance, *12: dulnesse, howsoeuer, euery, shortnesse* appears to consist of transitional or filler terms that do not suggest a distinct subject. Topic *19: heretic, punish, tongues, offence* contains strong terms that hint at a narrative but lack coherence. 'Punish' and 'offence' suggest crime and justice, while 'heretic' indicates deviation from accepted beliefs, and 'tongues' implies bad langauge. Together, they touch on challenging authority, yet they remain loosely connected. Similarly, *26: ought, virtue, authority, judiciary* includes terms that hint

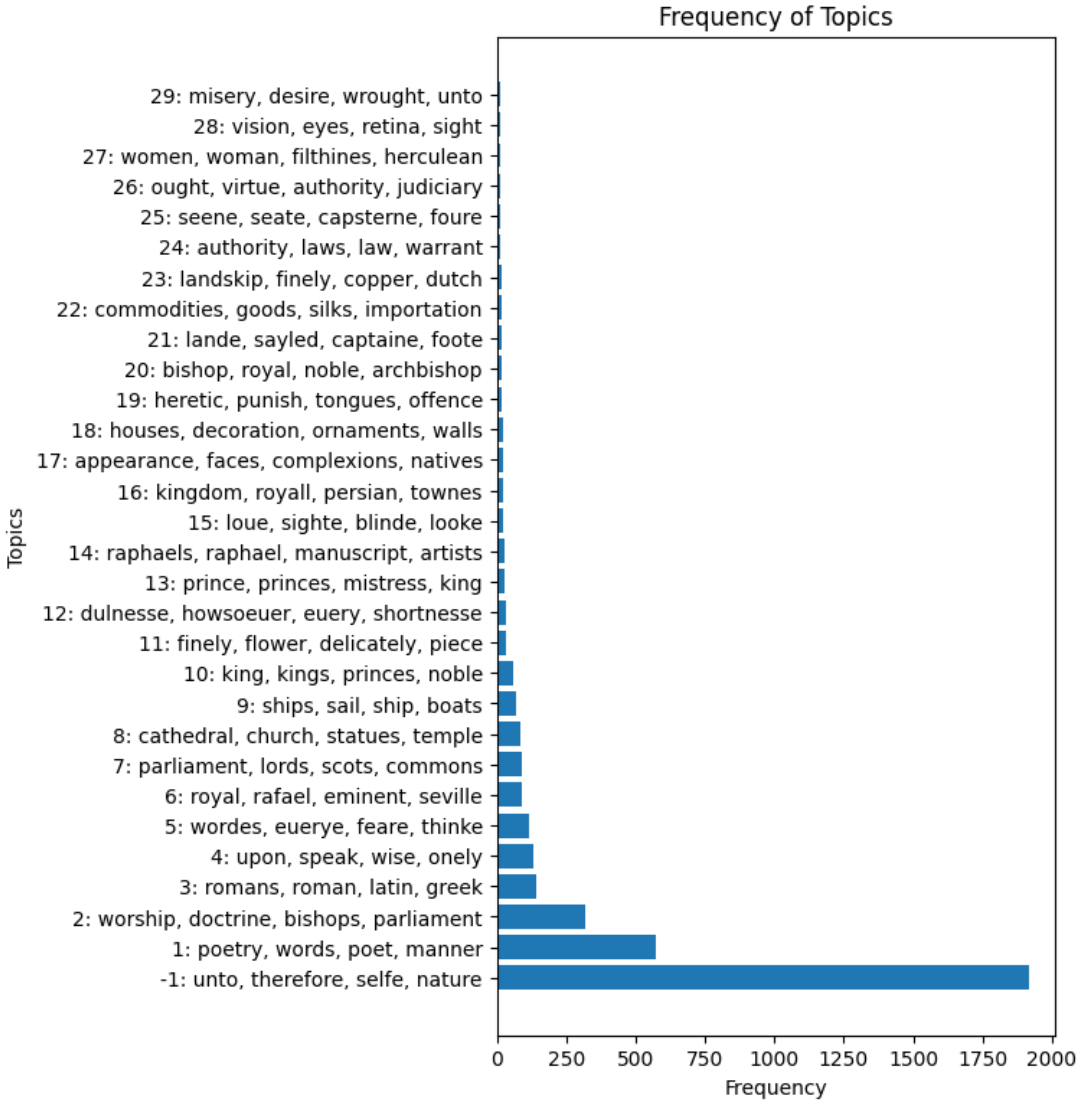


Figure 5.1: Frequency of the topics over the paragraphs.

at ethical or legal discussions but fail to form a clear and unified topic. These topics lack interpretability because they do not reflect one theme or subject that can be easily understood within the context of the dataset.

In Figure 5.1, the frequency of each topic is shown. The most striking feature of the distribution is the high frequency of topic *-1 unto, therefore, selfe, nature*, which stands significantly higher than all other topics, making the residual topic dominant in the corpus. The large size of this category suggests that numerous paragraphs are either general or do not align with the more specific topics related to the visual and conceptual aspects of art. The overwhelming size could indicate that many paragraphs, while thematically connected to broader art-related discussions, do not fit neatly into more focused, discrete categories. This may point to paragraphs that are disjointed, rather than strictly tied to artistic techniques or specific movements.

Normally, a dominant residual topic indicates a bad topic model or a bad dataset. However, due to the methodology of retrieving the art-related paragraphs, this result

is not unexpected. Since all paragraphs that contained certain art-related words were retrieved, a lot of general paragraphs are present in the dataset. Nonetheless, the topic model did find topics that are interpretable and expected in the art of that period, indicating a good result. Especially since the topics *1: poetry, words, poet, manner* and *2: worship, doctrine, bishops, parliament* are most frequent, which reflects the period of 1550 to 1800 very well. Topic 1 highlights the significant overlap between visual art and literature in the time when poetic works often inspired artistic creations, especially in the context of historical or mythological representations. Moreover, poetry was regarded as a distinct art form, and the term "poetic" was also used to describe paintings (Golahny, 2002). Topic 2 underscores the influential role that religious and political factors played in art during this time. The frequent use of terms related to "worship" and "doctrine" reflects the strong presence of religious themes, particularly the word "bishops" within the Catholic and Protestant contexts. Additionally, the mention of "bishops" and "parliament" indicates how art was shaped by the authorities of both religion and state, as commissions often originated from these institutions, reflecting their influence and ideologies in art. Art during this period was predominantly created for the elite rather than the common people (Lytle and Orgel, 1981). Because these topics are interpretable and accurately describe the period, I decided that I could utilize the topic model's results and disregard the residual topic.

5.1.2 Trend Analyse

In this section, I will analyze the results as a typical user of this framework, demonstrating the outcome of the trend analysis for naive users.

For each topic, I have created two visualizations: one showing the frequency and relative occurrence per decade, and another displaying sentiment relative to the average sentiment of that decade. The results and possible insights vary considerably across topics due to differences in data availability, distribution over time, and topic clarity. Many topics suffer from sparse data, irregular spikes in certain decades, or a highly uneven temporal distribution. This makes it difficult to derive consistent or meaningful trends, as the data is often too limited for statistical interpretation or skewed by a few decades of higher occurrence with little broader relevance.

In this section, I will therefore focus on a selected subset of topics that offer more interpretable patterns. Specifically, I discuss the topics *1: poetry, words, poet, manner*, *2: worship, doctrine, bishops, parliament*, *3: romans, roman, latin, greek*, *6: royal, rafael, eminent, seville*, *7: parliament, lords, scots, commons*, *8: cathedral, church, statues, temple*, and *9: ships, sail, ship, boats*. The complete set of topic trend visualizations can be found in Appendix A.

To maintain transparency in the output of the framework, I chose not to merge topics that appear closely related, even when such merging might improve interpretability. Instead, I present the topics exactly as identified by the model. Table 5.2 categorizes the topics according to data-related issues that affect their interpretability. For example, Topic 11 (finely, flower, delicately, piece) contains data from only three decades (1640, 1680, and 1770), with a significant spike in 1680 and very low counts in the others. This makes it insufficient for drawing meaningful conclusions about long-term trends.

The distribution of *1: poetry, words, poet, manner*, shown in Figure 5.2, reveals a pronounced shift over time. In the early 17th century, occurrences are extremely sparse; however, there is a marked increase in the later decades, with counts reaching

Category	Description	Topics
A. Widely Distributed but Consistently Low Counts	The topic appears in many decades, but the counts remain very low across those decades, so even though temporal coverage is good, the data are too weak to support reliable trends.	10: king, kings, princes, noble 12: dulnesse, howsoeuer, euery, shortnesse 13: prince, princes, mistress, king 14: raphaels, raphael, manuscript, artists 16: kingdom, royall, persian, townes 17: appearance, faces, complexions, natives 18: houses, decoration, ornaments, walls 19: heretic, punish, tongues, offence 20: bishop, royal, noble, archbishop 22: commodities, goods, silks, importation 24: authority, laws, law, warrant 25: seene, seate, capsterne, foure 27: women, woman, filthines, herculean 28: vision, eyes, retina, sight
B. Limited Decade Coverage	The topic is only recorded in a few decades, so the overall temporal span is too limited for any reliable trend analysis.	21: lande, sayled, captaine, foote 26: ought, virtue, authority, judiciary 29: misery, desire, wrought, unto
C. Data Concentrated in Only One/Few Decades	The topic appears in only one or a very few decades (or one decade dominates the counts), so no trend over time can be reliably assessed.	11: finely, flower, delicately, piece 23: landskip, finely, copper, dutch
D. Not interpretable topics	The topic shows a positive trend and has sufficient data, but is not relevant or interpretable for Art Historians.	4: upon, speak, wise, lonely 5: wordes, euerye, feared, thinke

Table 5.2: Topics categorized per issue for trend analyses.

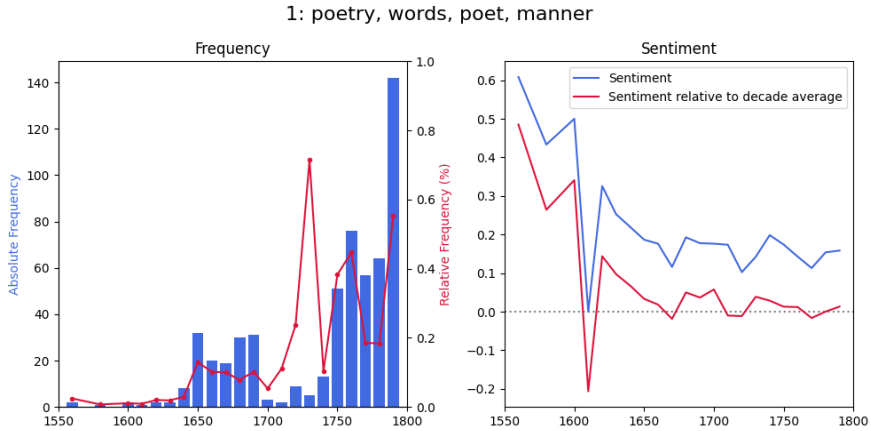


Figure 5.2: Trend analysis of topic *1: poetry, words, poet, manner.*

76 in 1760 and 142 in 1790. The corresponding rise in relative occurrence indicates an emerging prominence of poetic discourse in the latter part of the period. This temporal dynamic may reflect evolving literary interests or changes in the publishing landscape that favored poetic expression. The sentiment plot reveals an interesting trend: while early decades exhibit muted sentiment (in line with the very low frequency), the later decades show a noticeable shift toward a more positive or enthusiastic tone. This change corresponds with the increased frequency of these terms, implying that as poetic discourse became more prominent, its associated emotional valence also shifted, potentially reflecting evolving aesthetic standards or a growing appreciation for poetic forms.

Topic *2: worship, doctrine, bishops, parliament* is well represented across multiple decades, though with notable variability. Early in the period, the frequencies are minimal (with some decades recording only a single occurrence), yet later periods show moderate usage with relative occurrences clustering between 0.04 and 0.18. The persistence of these terms throughout the timeline indicates a continued engagement with religious and political discourse. Such a pattern may point to the evolving nature of institutional debates and doctrinal conflicts during the period, suggesting that these concepts retained a fluctuating yet enduring relevance in the textual corpus. The sentiment profile shows moderate variability, with slight peaks that align with known periods of religious and political turbulence. Although the overall sentiment does not veer into extremes, the nuanced fluctuations may be indicative of moments when debates over doctrine or ecclesiastical authority were particularly charged. This layered reading suggests that while the usage remained relatively constant, the tone with which these terms were employed was sensitive to contemporary debates.

The temporal pattern for Topic *3: romans, roman, latin, greek* shows that there was an overall low frequency of references in the earlier and middle periods, suggesting that the Renaissance was not particularly influential in England during this time. However, there was a significant increase in references in the late 18th century, with counts of 49 in 1780 and 37 in 1770. This uneven distribution implies that classical languages and associated cultural references were generally peripheral for much of this period, but they became notably more prominent toward its end. The sentiment plot is particularly intriguing when contrasted with its frequency. For much of the period, the emotional tone remains subdued, which is consistent with its low frequency. However,

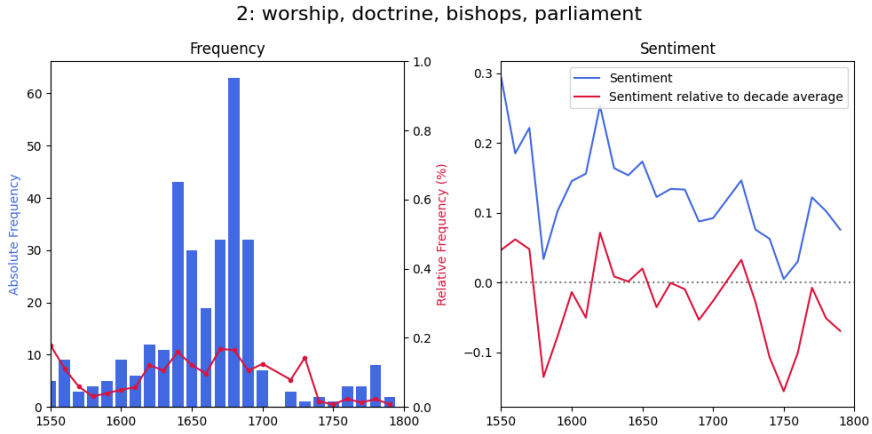


Figure 5.3: Trend analysis of topic *2: worship, doctrine, bishops, parliament.*

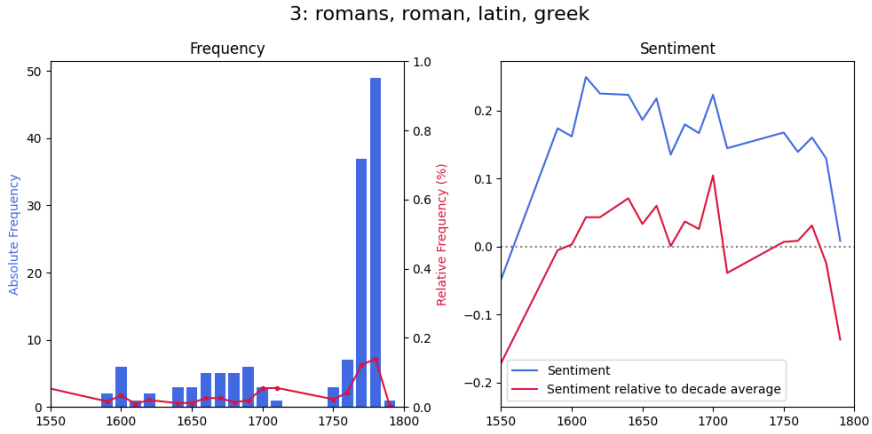


Figure 5.4: Trend analysis of topic *3: romans, roman, latin, greek.*

the pronounced frequency spike in the late 18th century is accompanied by a marked shift in sentiment, suggesting that references to classical languages and culture were suddenly imbued with a different emotional quality. The late surge may reflect the impact of the Enlightenment-era classical revival or changing intellectual trends that revalued antiquity.

The data for topic *6: royal, rafael, eminent, seville* are dominated by a single, substantial spike in 1780, where the frequency reaches 85, contrasting sharply with near-zero counts in other decades. This isolated surge suggests that the thematic elements of royalty and eminent figures, possibly including localized references such as "Seville," experienced a brief but intense period of attention. The sentiment data for the topic reveals an intriguing pattern as well. The single, substantial spike in frequency during 1780 is accompanied by an elevated emotional tone. This heightened sentiment during the spike suggests that the usage of these terms in that decade was not merely a quantitative aberration but was also qualitatively distinct. The outlier may be driven by a specific cultural or political event, or by a concentration of texts from that particular decade, and thus should be interpreted as a localized phenomenon rather than as evidence of a sustained trend. However, this does align with the trend seen in topic 3, since *rafael* and *seville* are closely related to *roman.*

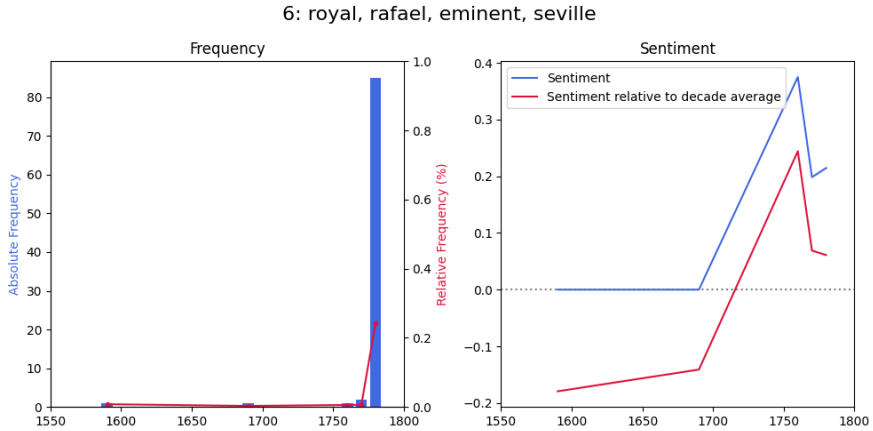


Figure 5.5: Trend analysis of topic *6: royal, rafael, eminent, seville.*

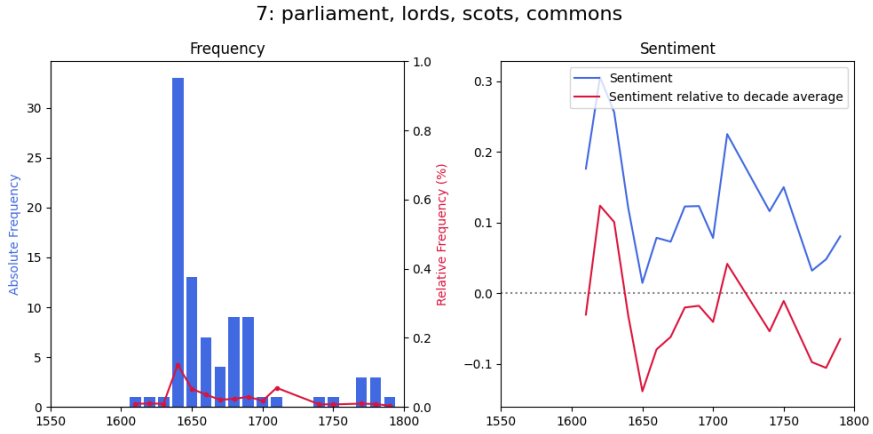


Figure 5.6: Trend analysis of topic *7: parliament, lords, scots, commons.*

Although topic *7: parliament, lords, scots, commons* is present in multiple decades, its overall frequency remains consistently low. The sustained low counts across the timeline indicate that while the subject of parliamentary and political institutions is acknowledged in the corpus, it never emerges as a dominant theme. This could suggest that references to such institutional matters were either confined to specific contexts or served a supplementary role within the broader discourse. Consequently, the data imply a peripheral yet stable engagement with political terminology that does not rise to the level of a major trend.

Topic *8: cathedral, church, statues, temple* does present an anomaly in 1800, where the relative occurrence spikes to 1.0. This value is attributable to the size of only 1 for that particular decade. Excluding this outlier, the data shows that there are relatively few instances of this topic. What is particularly interesting is the low relative coverage of these subjects in the earlier decades, when religion was the most prominent theme in art, suggesting that this topic should have been more widely represented. This is quite likely due to the preprocessing step, where highly religious texts were excluded from the data. Additionally, the peak in 1740 may indicate the onset of the Industrial Revolution and its counter-movement, Romanticism. The sentiment analysis generally aligns with its moderate frequency distribution; however, there is an

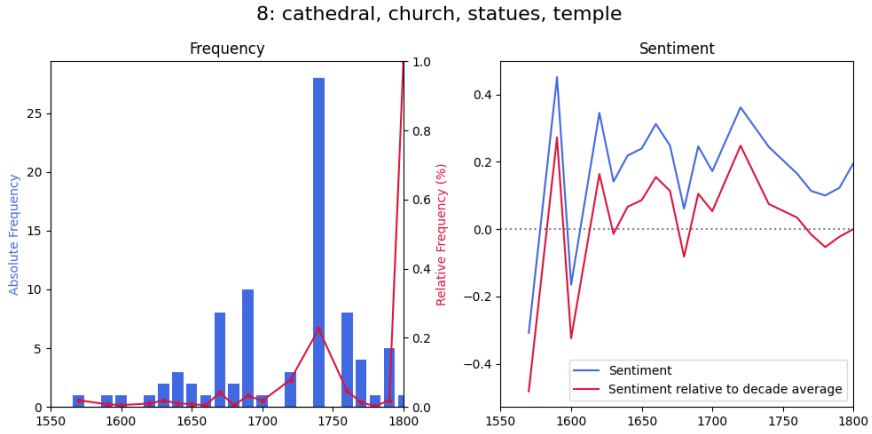


Figure 5.7: Trend analysis of topic *8: cathedral, church, statues, temple.*

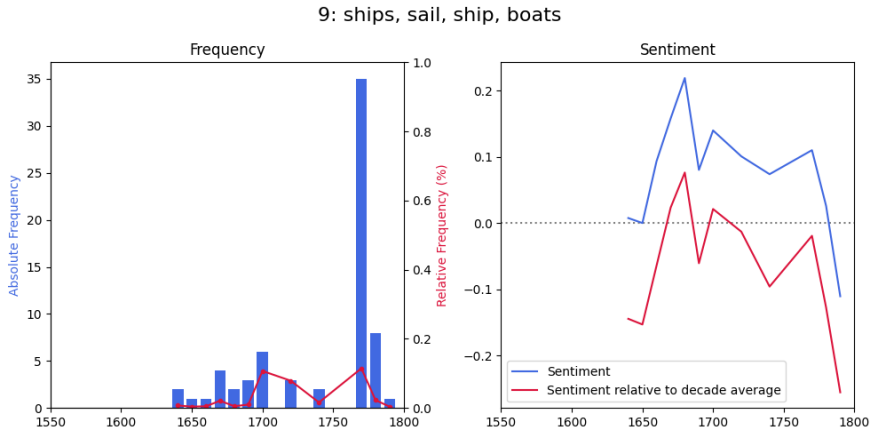


Figure 5.8: Trend analysis of topic *9: ships, sail, ship, boats.*

outlier in the 1800 data point where sentiment appears anomalously polarized. This spike mirrors the frequency anomaly in 1800 and likely reflects a small-sample artifact rather than a substantive shift in sentiment. Overall, aside from this isolated deviation, the sentiment remains balanced, consistent with the enduring symbolic significance of religious imagery in the corpus.

Topic *9: ships, sail, ship, boats*: shows a variable pattern, with some decades registering minimal counts while others indicate moderate frequencies. This inconsistency suggests that maritime terminology was subject to episodic emphasis rather than a continuous narrative. The fluctuations could be attributed to historical events affecting maritime trade or naval conflicts, or to shifts in the types of texts being published. The sentiment plot shows occasional peaks and troughs that appear to correspond with specific historical events or shifts in maritime discourse. However, the overall ambiguous pattern cautions against drawing firm conclusions about a sustained maritime affective trend.

5.1.3 Evaluation of the models

To validate the naive users conclusions, the models were evaluated using automatic metrics, a survey, and error analysis.

	Latent Topic	Intruder	Score
-1	**unto, therefore, selfe, nature**	**tongues**	**0.1667**
1	poetry, words, poet, manner	copper	1.0000
2	**worship, doctrine, bishops, parliament**	**seene**	**0.5833**
3	romans, roman, latin, greek	judiciary	0.9167
4	**upon, speak, wise, onely**	**laws**	**0.1667**
5	wordes, euerye, feare, thinke	bishop	0.8333
6	**royal, rafael, eminent, seville**	**eyes**	**0.3333**
7	parliament, lords, scots, commons	latin	0.8333
8	cathedral, church, statues, temple	sighte	1.0000
9	ships, sail, ship, boats	dulnesse	1.0000
10	king, kings, princes, noble	decoration	1.0000
11	finely, flower, delicately, piece	therefore	0.8333
12	**dulnesse, howsoeuer, euery, shortnesse**	**mistress**	**0.4167**
13	prince, princes, mistress, king	thinke	1.0000
14	**raphaels, raphael, manuscript, artists**	**wordes**	**0.5000**
15	**loue, sighte, blinde, looke**	**onely**	**0.5000**
16	**kingdom, royall, persian, townes**	**speak**	**0.6667**
17	appearance, faces, complexions, natives	unto	0.9167
18	houses, decoration, ornaments, walls	rafael	0.9167
19	**heretic, punish, tongues, offence**	**euerye**	**0.4167**
20	bishop, royal, noble, archbishop	ship	1.0000
21	lande, sayled, captaine, foote	rafael	0.9167
22	commodities, goods, silks, importation	euerye	1.0000
23	**landskip, finely, copper, dutch**	**sight**	**0.1667**
24	authority, laws, law, warrant	ships	0.7500
25	**seene, seate, capsterne, foure**	**piece**	**0.1667**
26	**ought, virtue, authority, judiciary**	**euerye**	**0.6667**
27	**women, woman, filthines, herculean**	**cathedral**	**0.0833**
28	vision, eyes, retina, sight	princes	1.0000
29	**misery, desire, wrought, unto**	**ought**	**0.0833**

Table 5.3: Topic Intrusion Survey results per topic, with the intruder word and the percentage of votes for the intruder. Bold topics are considered not to be interpretable by the crowd, since they got a score less than 0.75.

Topics

Topic results were evaluated using an intrusion survey for coherence and an error analysis on topic assignment.

Intrusion Survey The results of the Topic Intrusion Survey are presented in Table 5.3. The table displays the latent topics, the identified intruder words, and the corresponding intrusion scores, which indicate the proportion of annotators who correctly identified the intruding word.

Topics that received an intrusion score of less than 0.75 were considered non-interpretable, as their intruder words were not consistently identified by the annotators. These topics are highlighted in bold in the table. Out of the 30 topics evaluated, 13 were determined to be non-interpretable based on this threshold. The intrusion score threshold of 0.75 was chosen to reflect statistically significant consensus among annotators. A binomial test shows that 3 out of 4 (0.75%) correct responses yields a p-value of 0.027 (p ¡ 0.05). In contrast, a score of 0.667 (e.g., 2 out of 3 correct) has a p-value of 0.10 and can therefor not be distinguished from random guessing.

Several topics exhibited high interpretability, with intrusion scores of 1.00, indicating perfect agreement among annotators in recognizing the intruder word. Examples of such topics include *poetry, words, poet, manner* with the intruder "copper", and *cathedral, church, statues, temple* with the intruder "sighte," and *ships, sail, ship, boats* with the intruder "dulnesse." These results suggest that the words within these topics form coherent themes, making the intruder word easily distinguishable.

Conversely, topics such as *unto, therefore, selfe, nature* with the intruder "tongues" (0.1667) and *landskip, finely, copper, dutch* with the intruder "sight" (0.1667) had low scores, indicating that annotators struggled to identify the out-of-place term. This suggests that these topics lack clear thematic cohesion, leading to difficulties in interpretation.

Overall, the topic intrusion evaluation highlights the varying degrees of coherence among the extracted topics. The presence of several non-interpretable topics suggests potential improvements in the topic modeling approach, such as refining the number of topics or adjusting model parameters to enhance topic distinctiveness and interpretability.

Error Analysis on Topic Alignment To assess the quality of the model's topic assignments, a manual evaluation was conducted in which I attempted to select the appropriate topic for the development set, reproducing the survey intended to be conducted. The model's original topic assignments were taken as the reference standard, and any deviation in the researcher's selection was treated as an error. The overall accuracy of these manual assignments was 0.383, indicating a substantial level of disagreement with the model. While this may suggest low alignment between human interpretation and model inference, it also provides insight into where and why the model's topic decisions may be difficult to recover or justify from a human perspective.

A topic-wise breakdown, shown in Table 5.4, reveals that certain model-assigned topics were more intuitively recoverable than others. For example, topics such as *unto, therefore, selfe, nature* and *poetry, words, poet, manner* were frequent (n = 26 and n = 16, respectively) and had moderate alignment rates of 0.385 and 0.375. In contrast, several topics with perfect accuracy, such as *finely, flower, delicately, piece*, *landskip,*

Table 5.4: Topic-wise Accuracy and Counts

Correct Topic	**Accuracy**	**n**
finely, flower, delicately, piece	1.000	1
landskip, finely, copper, dutch	1.000	3
loue, sighte, blinde, looke	1.000	1
ships, sail, ship, boats	1.000	1
women, woman, filthines, herculean	1.000	1
unto, therefore, selfe, nature	0.385	26
poetry, words, poet, manner	0.375	16
cathedral, church, statues, temple	0.000	2
dulnesse, howsoeuer, euery, shortnesse	0.000	1
heretic, punish, tongues, offence	0.000	1
raphaels, raphael, manuscript, artists	0.000	2
romans, roman, latin, greek	0.000	1
royal, rafael, eminent, seville	0.000	2
wordes, euerye, feare, thinke	0.000	2

finely, copper, dutch, and *ships, sail, ship, boats*, were assigned correctly in all cases, but occurred only once or a few times.

Some topics were consistently difficult for the researcher to identify, including *cathedral, church, statues, temple* and *romans, roman, latin, greek*, each with zero correct selections. Since the findings from the topic intrusion survey indicate that these topics are interpretable, the wrong alignment selection could suggest a mismatch between the model's understanding of a topic and how humans interpret it. Alternatively, it might reflect noise and ambiguity in the way the topics appear in the source texts.

A temporal analysis shows considerable variation in error rates across decades. Error rates of each decade are visible in (Table 5.5). Very early texts (e.g., from the 1580s, 1610s, and 1630s) had 1.00 error rates, likely reflecting linguistic and semantic shifts that made the model's topics less accessible to contemporary human reasoning. More recent decades, such as the 1680s through 1790s, generally performed better, though errors remained substantial (e.g., 0.75 in the 1780s and 0.636 in the 1790s). This trend supports the interpretation that the model's topics, derived from modern English corpora, become increasingly intuitive as the language of the text approaches contemporary usage.

Overall, this error analysis highlights not only which topics are more or less human-recoverable but also how both topic clarity and linguistic familiarity affect performance. It also shows the difficulty of the task itself.

Sentiment

The analysis of the relationship between the mean ratings provided by human annotators for the sentiment survey and the predictions made by the model was measured using Spearman's rank correlation and Root Mean Squared Error. In Table 5.6, the number of annotators per section is visible. The discrete nature of human ratings, which can range from -3 to +3, introduces some potential noise in the analysis. However, the primary goal of this study is to rigorously evaluate the accuracy of the model. Since the model's performance is being compared to human ratings, it is essential to

Table 5.5: Error Rate by Decade

Decade	Error Rate
1560	0.500
1580	1.000
1610	1.000
1620	0.500
1630	1.000
1640	0.000
1650	0.500
1660	0.750
1670	0.667
1680	0.375
1690	1.000
1720	0.000
1760	0.500
1770	0.800
1780	0.750
1790	0.636

use a significance threshold of 0.05. This threshold will help ensure that any observed relationships are statistically reliable. Therefore, I will use 0.05 to determine whether to reject the null hypothesis, which states that there is no significant monotonic relationship between human ratings and model predictions. In other words, this implies that the rankings of the two variables are independent of each other.

The results showed a weak positive correlation of 0.231. This indicates a slight tendency for both the human annotators' ratings and the model's predictions to increase together. However, the correlation is not statistically significant, as the p-value of 0.076 exceeds the conventional threshold of 0.05. Therefore, I fail to reject the null hypothesis, which states that there is no significant monotonic relationship between the two sets of scores.

Additionally, the model's Root Mean Squared Error (RMSE) was found to be 0.39, indicating that the average deviation between the human ratings and the model's predictions is approximately 0.39 units on a scale from -1 to 1. While this suggests that the model is reasonably close to human ratings, it highlights room for improvement in the model's accuracy.

In conclusion, the results suggest that while there is a slight positive correlation between the model's predictions and the human ratings, this relationship is not statistically significant. Furthermore, the RMSE of 0.39 indicates that the model's predictions are somewhat accurate but still exhibit noticeable errors. Therefore, the model's predictions do not reliably reflect the human annotators' judgments, and further im-

Section	1	2	3	4	5	6
Annotators	13	2	2	1	3	1

Table 5.6: Number of annotators per section for the Sentiment Survey.

provements are necessary to achieve better alignment. However, there is enough alignment to use the results for this exploratory study, since the primary aim is to identify general patterns and assess the feasibility of computational sentiment analysis in historical texts, rather than to provide definitive answers. Additionally, the model was not adapted or fine-tuned for this specific domain, so the observed partial alignment is promising as a baseline. However, this misaligned needs attention in further work, and is taken into account by making any conclusions.

Error Analysis by Decade To better understand the temporal performance of the sentiment analysis model, I evaluated its predictions against the human-annotated sentiment ratings across the decades. For each decade, the Relative Mean Absolute Error (RMAE) and Spearman's rank correlation coefficient between the model's predictions and the scaled mean human ratings are reported. Additionally, I examined the mean sentiment of each decade and the mean prediction error, the latter indicating whether the model systematically overestimated or underestimated sentiment. Results of this error analysis are presented in Table 5.7

Table 5.7: Sentiment Analysis Performance by Decade

Decade	n	**RMAE**	**Spearman**	**Mean Sentiment**	**Mean Prediction Error**
1560	2	0.625	1.000	0.111	0.139
1620	2	2.904	-1.000	0.154	-0.346
1630	2	1.125	-1.000	0.333	-0.125
1650	2	0.939	1.000	0.000	0.082
1660	4	1.089	0.000	0.042	0.179
1670	3	8.106	-0.866	-0.043	0.346
1680	8	0.936	0.417	0.184	0.051
1760	8	1.152	0.168	0.068	0.246
1770	5	0.711	0.103	0.411	-0.257
1780	8	1.059	0.146	0.063	0.027
1790	11	0.822	0.548	0.246	-0.059

While some decades (such as the 1560s, 1620s, 1630s, and 1650s) display extreme Spearman correlation coefficients of +1.000 or -1.000, these results are based on only two samples each. With such low sample sizes, these correlations are not robust and should be interpreted cautiously. Examining the underlying data shows that these extreme values are simply mathematical artifacts: with only two data points, any monotonic relationship, whether both model and human ratings increase, decrease, or move in opposite directions, will always yield a perfect positive or negative correlation. These outcomes do not reflect meaningful trends or model performance, but rather the inherent sensitivity of correlation statistics to very small sample sizes. Therefore, no substantive conclusions can be drawn for these early decades.

More meaningful patterns emerge in decades with larger sample sizes, particularly the 1680s, 1760s, 1770s, 1780s, and 1790s, each containing between 5 and 11 annotated samples. In these periods, the Spearman correlation coefficients are consistently positive but relatively modest (e.g., 0.417 in the 1680s, 0.168 in the 1760s, 0.548 in the 1790s), indicating that the model captures some ordinal structure in human sentiment ratings but does so only partially.

The 1680s stand out for their relatively low RMAE (0.936) and low mean prediction error (0.051), suggesting both accurate and unbiased predictions. In contrast, the 1760s and 1770s show higher RMAE values (1.152 and 0.711, respectively), accompanied by consistent under- or over-predictions: the model over-predicts sentiment in the 1760s (mean error = +0.246), while it under-predicts in the 1770s (mean error = –0.257). This might reflect shifting vocabulary or affective connotations that the model does not adequately adjust for.

Interestingly, the 1790s, the decade with the largest number of annotated texts (n = 11), show moderate accuracy (RMAE = 0.822) and the highest Spearman correlation (0.548) among the decades with sufficient data, indicating relatively reliable performance. However, the average error remains slightly negative (–0.059), suggesting a minor tendency to underestimate sentiment in this period.

The 1670s present a particularly challenging case. With only three samples, the model's predictions (0.47, 0.16, 0.28) show consistently positive sentiment, whereas the human means (-0.14, 0.00, 0.00) are neutral or slightly negative. This mismatch results in both the highest RMAE (8.106) and a strong negative correlation (–0.866) for this decade. However, with such a small sample size, both the error and correlation metrics are highly sensitive to individual data points and may not reliably reflect true model performance. Therefore, these results should be interpreted with caution as well.

Overall, the sentiment model's performance varies across decades, with some periods showing moderate predictive accuracy and others marked by notable errors. While decades with very few data points (e.g., 1560, 1620, 1630) show extreme Spearman correlations, these results are unreliable due to small sample sizes. More stable periods, such as the 1670s–1790s, reveal systematic biases: the model tends to underestimate sentiment in some decades (e.g., 1770s, 1790s) and overestimate it in others (e.g., 1760s). These patterns likely stem from the fact that the sentiment model was trained on modern English texts, whereas the input data reflects early modern language.

Conclusion

It is clear from the evaluation that the models are not fully capable of the desired results, and therefore the conclusions made by the naive users may not be completely trustworthy. However, the topics discussed were looked into, and most trends seem right. Therefore, even though the framework is not fully capable of capturing trends, some topics were promising.

5.2 Use case: Landscape paintings

Landscape painting in England evolved from decorative backdrops to symbols of national identity (Budnick, 2017), providing an interesting use case for this study. I employ the alternative framework, which uses keyword expansion and fuzzy matching to retrieve all landscape-related paragraphs for trend analyse with term-frequency and sentiment analysis, instead of topic modelling.

The first step in the alternative framework is query expansion. I began with the word 'landscape' and received the following expansions:

Query: scenery, topography, landscapes

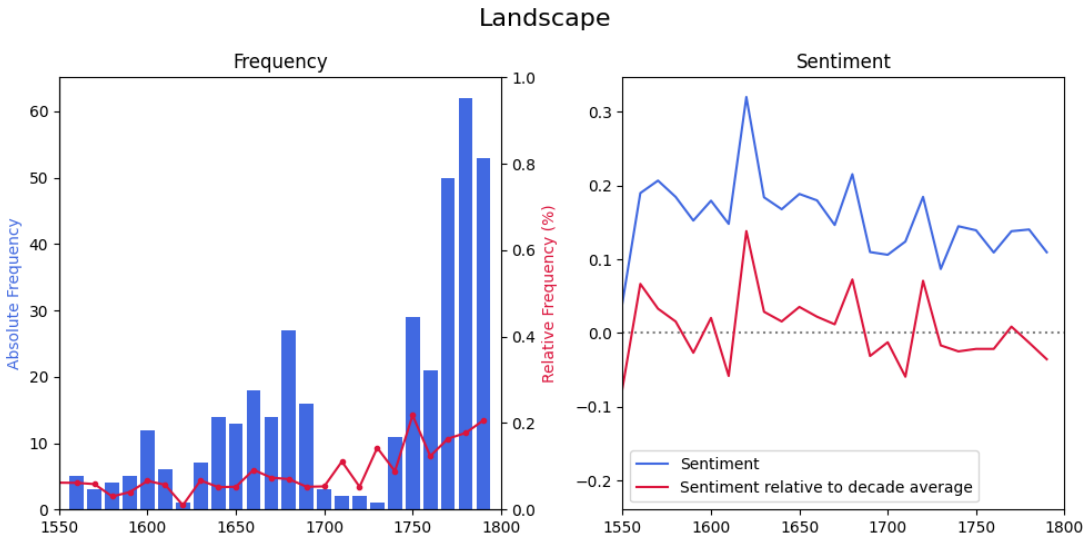


Figure 5.9: Trend analysis of landscape-related paragraphs.

For each of these, I looked up different spellings used between 1550 and 1800 in the Oxford English Dictionary. These results are visible in Table 5.8. Since 'landscape' and 'landscapes' appear in the same dictionary entry, I did not use "landscapes" further.

This resulted in a total of 380 paragraphs out of the 3974. The trends are visible in Figure 5.9.

The absolute frequency of landscape-related terms remained relatively low and stable between 1550 and 1650, with only minor fluctuations. However, it began to decrease around 1700 and then rose sharply between 1750 and 1800. The relative frequency shows a similar pattern, indicating that landscapes became a more significant topic in relation to the overall content. The sentiment analysis reveals mostly positive fluctuations, with a peak occurring between 1600 and 1650. Additionally, the sentiment relative to the average for each decade highlights periods when landscape discourse was framed more or less positively compared to general sentiment trends.

The frequency patterns closely align with the historical development of landscape painting in England. The low frequency in the 16th and early 17th centuries reflects the landscape's status as a secondary genre, often incorporated into religious or portrait paintings instead of being treated as an independent subject (Levy, 1974). A

Query	**Spelling variations**
scenery	scenerie, scenary
landscape	landtschap, lantschape, landt-shape, landscap, landskap, landskape, landschape, landshape, landchape, landscape, landskip, lantskip, landt-skip, lantscip, lantschip, lanscippe, landskipp, lantskipp
topography	topographye, topographie

Table 5.8: Spelling variations of the query words according to Oxford English Dictionary (Oxford English Dictionary, 1857)

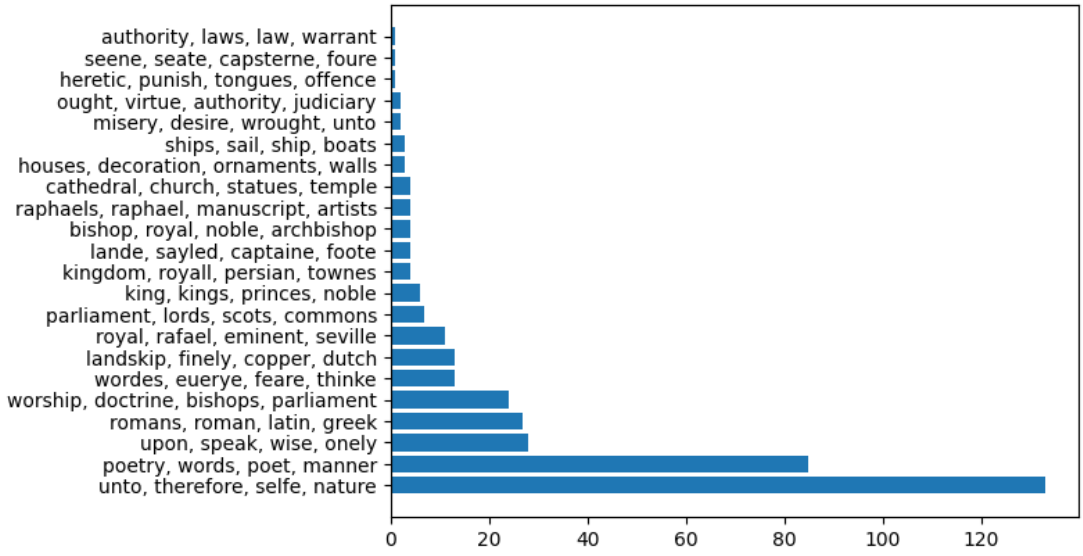


Figure 5.10: Distribution of derived landscape paragraphs across topics within the general framework.

slight increase in frequency during the late 17th century coincides with the influence of Dutch and Flemish landscape traditions, which gradually gained recognition in England (Karst, 2021).

After 1750, there is a sharp rise in both absolute and relative frequency, coinciding with the growing prestige of landscape painting. This connection between textual discourse and artistic developments highlights a broader transformation in aesthetic appreciation. By the late 18th century, landscape painting reached new heights of recognition, paving the way for the Romantic movement (Budnick, 2017; Hunt, 1985).

However, the sentiment does not seem to reflect these historical events accurately. The peak between 1600 and 1650 appears unusual, and the overall trend is stable—or perhaps even slightly decreasing. Art historians argue that sentiment towards landscapes should have changed during the Romantic movement at the end of the 18th century, when nature was idealized and appreciated. Therefore, it would have been more logical for sentiment to increase during this period.

In Figure 5.10, an analysis of the distribution of paragraphs across various topics derived from the general framework is shown. It is noteworthy that the majority of paragraphs are categorized under the residual topic. This observation is quite logical, as the topic features the term 'nature.' The disproportionate presence of paragraphs within this topic underscores the critical necessity of utilizing the specific framework for a more comprehensive examination. This second framework is specifically designed to delve deeper into the trends and nuances that the general framework may not fully reveal.

The trend of the landscape use-case is in line with what is expected from literature, and therefore this alternative framework is promising. By leveraging this framework, more robust insights and patterns are uncovered that contribute to a richer understanding of the trend at hand.

Chapter 6

Discussion

This chapter discusses the key limitations of this study and suggests future improvements. The main challenges identified include the evaluation of sentiment analysis and topic modeling, dataset limitations, technical constraints such as OCR errors, and the applicability of computational methods to historical art discourse.

Evaluation Challenges

A significant limitation of this study is the evaluation of both sentiment analysis and topic modeling based on survey responses. The obtained scores, while close to acceptable thresholds, were not high enough to ensure fully reliable results. The limited number of responses (12 and 14 participants respectivily) makes it difficult to assess the true statistical reliability of these evaluations.

In particular, the low agreement in the topic intrusion survey suggests that some topics may lack clear interpretability, indicating the need for further validation. Increasing the number of annotators, particularly those with domain expertise in art history, could improve reliability. Additionally, refining annotation guidelines, such as providing detailed examples and clearer criteria for topic categorization, could help ensure more consistent evaluations, as annotators currently rely solely on an online form for guidance.

An additional limitation concerns the interpretability of model-assigned topics, as assessed through a manual topic alignment task in which I attempted to recover the model's topic labels for the development set. The overall alignment accuracy was 0.383, indicating substantial difficulty in mapping the model's topics to human judgment.

Dataset Limitations

Another limitation is the presence of gaps in the dataset. The initial filtering of texts using art-related keywords may have unintentionally excluded relevant discussions that used different terms. Artistic trends are often implied through context rather than clearly defined by fixed vocabulary. As a result, some subtle but important shifts in artistic discourse may have been missed. Additionally, while the preprocessing step of removing highly religious texts is well motivated, it also affects the reliability of trends in religious topics.

A more robust text retrieval strategy incorporating semantic similarity measures, such as word embeddings or transformer-based retrieval methods, could address this issue. For instance, applying contextual word representations like BERT or Word2Vec

could capture variations in terminology and improve the comprehensiveness of the dataset.

Furthermore, the limited number of extracted paragraphs from the period between 1700 and 1730 presents challenges. This scarcity of data may distort trend analysis, making certain artistic developments appear less significant than they actually were. Some artistic shifts might be underrepresented or misinterpreted due to the limited availability of digitized sources from that timeframe. Future research could mitigate this issue by incorporating additional historical archives or adjusting data selection criteria to reduce temporal imbalances.

Technical Constraints

Errors introduced by Optical Character Recognition (OCR) present another challenge. The texts from ECCO contain OCR errors that introduce noise, affecting both topic modeling and sentiment analysis.

Spelling variations and archaic language further complicate processing, as they can lead to incorrect tokenization or misclassification of terms. For example, historical spelling differences (e.g., "landskip" instead of "landscape") can result in incorrect parsing by modern NLP models. Some efforts were taken to reduce these problems, the manual step of searching for the different spelling variations in the second framework and the use of the Levenshtein distance. However, these problems remained for the topic model and the sentiment analysis. Addressing these issues could involve training OCR models specifically for historical texts or incorporating human verification for high-confidence errors.

Another potential solution is the use of lexicon expansion methods tailored for historical English. For instance, domain-specific spelling normalization tools could improve tokenization accuracy and reduce errors in downstream analyses.

BERTopic and Sentiment Analysis Model Limitations

BERTopic, while providing a context-aware approach to topic modeling, has not been widely studied for historical texts. The maximum coherence score of 0.492 suggests that some topics may lack clear interpretability. The topic intrusion survey of the extracted topics indicates that 13 out of the 29 topics were not interpretable. This raises the question of whether alternative modeling approaches, such as dynamic topic modeling or hierarchical topic modeling, could yield better results. The Latent Dirichlet Allocation method was also tested, but it produced worse results.

Similarly, TextBlob, used for sentiment analysis, relies on a general-purpose lexicon that is not optimized for historical texts. As a result, certain words underwent semantic shifts that led to misclassification. A domain-adapted sentiment analysis model, trained on historical language conventions, would likely yield more accurate results. Specifically, the sentiment analysis achieved a p-value of 0.076 alignment with human annotators, meaning there is no significant monotonic relationship between the annotators and the models scores.

Future improvements could involve fine-tuning sentiment analysis models on historical corpora or leveraging embeddings trained on Early Modern English and Modern English texts to capture period-specific language use.

By addressing these limitations, future research can enhance the integration of computational techniques in art history, ensuring that text-mining methods continue

to provide valuable insights into artistic discourse over time. However, these limitations are not easy to adress, since it is hard to get enough corpus of Early Modern English and Modern English texts to train such models. Therefore it might be wise to look at other models and tools as well.

Chapter 7

Conclusion

The aim of this study was to explore the possibilities of using text-mining techniques in art history research, by identifying trends. By applying BERTopic to a corpus of art-related texts from 1550 to 1800, distinct thematic structures were uncovered, providing insights into the evolution of artistic and cultural discussions. Additionally, sentiment analysis was employed to evaluate the emotional tone associated with these topics over time. The results offer a nuanced view of the interplay between art and literature, highlighting trends and their transformations across centuries.

The topic modeling did succeed in identifying several clusters that align with known historical themes, such as monarchy, religion, maritime activities, and classical influences. Which demonstrating some capacity of computational methods to capture meaningful patterns. However, a notable proportion of topics proved uninterpretable or overlapped considerably, reflecting the inherent complexity and ambiguity of historical texts. The presence of a large residual topic further suggests that the model struggled to account for the full diversity of the corpus, limiting the clarity and reliability of some findings.

The topic intrusion survey reinforced these concerns: while some topics were coherent and interpretable, nearly half were difficult for human annotators to distinguish. This points to ongoing challenges in achieving both accuracy and interpretability in unsupervised topic modeling, particularly with heterogeneous, context-rich historical data.

Sentiment analysis also yielded mixed results. Although there was a weak positive correlation with human ratings, this alignment was not robust, likely due to the limitations of applying contemporary sentiment models to early modern texts. This underlines the need for more historically tools and approaches when analyzing emotional tone in art historical sources.

Trend analysis revealed fluctuations in themes such as poetry, religious discourse, and classical influences, echoing shifts documented in the literature. However, some topics suffered from sparse or irregular data, constraining the reliability of certain trends and highlighting the risks of over-interpreting computational outputs.

Importantly, the application of an alternative analytical framework proved more successful. This approach not only uncovered the use case trend consistent with established art historical scholarship about landscape art, but also provided more robust and interpretable patterns, suggesting that careful methodological choices can substantially improve the quality of insights derived from text mining.

In summary, while this study demonstrates both the promise and the pitfalls of

integrating natural language processing into art history research, it also highlights the need for critical reflection and methodological refinement. Computational approaches can reveal patterns and developments that are difficult to discern manually, but their limitations, especially regarding interpretability and data coverage must not be overlooked.

Future research could build on this work by applying the second framework to other artistic movements, such as Romanticism or Impressionism, to test whether similar linguistic patterns emerge. Doing so would help evaluate the generalizability of this approach across different artistic and historical contexts.

In addition, expanding the dataset to include non-English texts would offer a more comprehensive view of art historical discourse. While this study focused on English-language material, significant contributions to art theory and criticism have appeared in French, German, and Italian sources. Incorporating multilingual topic modeling techniques could enable cross-cultural comparisons and enrich the analysis of how art has been discussed across time and geography.

Bibliography

A. Allaith, K. N. Degn, A. Conroy, B. S. Pedersen, J. Bjerring-Hansen, and D. Hershcovich. Sentiment classification of historical danish and norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands, 2023. University of Tartu Library.

R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

L. Azzopardi, M. Girolami, and K. Van Risjbergen. Investigating the relationship between language model perplexity and ir precision-recall measures. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 369–370, 2003.

S. Baccianella, A. Esuli, F. Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta, 2010.

J. Barnes, R. Klinger, and S. Schulte im Walde. Sentiment annotators are not a gold standard. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 52–62, 2021.

A. Behdenna, F. Barigou, and G. Belalem. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Informatics and Software Technology*, page 107492, 2024.

E. Biswas, M. E. Karabulut, L. Pollock, and K. Vijay-Shanker. Achieving reliable sentiment analysis in the software engineering domain using bert. In *2020 IEEE International conference on software maintenance and evolution (ICSME)*, pages 162–173. IEEE, 2020.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57 (2):1–30, 2010.

M. Bowman. Text-mining metadata: What can titles tell us of the history of modern and contemporary art? *Journal of Cultural Analytics*, 8(1), June 2023. doi: 10.22148/001c.74602. URL https://doi.org/10.22148/001c.74602.

B. L. Budnick. *The Lay of the Land: English Landscape Themes in Early Modern Painting in England*. PhD thesis, University of Washington, 2017.

Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji. A novel neural topic model and its supervised extension. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, 2009.

Q. Chen, M. El-Mennaoui, A. Fosset, A. Rebei, H. Cao, P. Bouscasse, C. E. O'Beirne, S. Shevchenko, and M. Rosenbaum. Towards mapping the contemporary art world with artlm: an art-specific nlp model, 2022. URL https://arxiv.org/abs/2212.07127.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

F. Diaz, B. Mitra, and N. Craswell. Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 367–377, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1035. URL http://aclweb.org/anthology/P16-1035.

A. B. Dieng, F. J. Ruiz, and D. M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Gale Cengage. Eighteenth Century Collections Online (ECCO), 2019. URL https://eebo.chadwyck.com. Accessed: February 27, 2024.

A. Golahny. *The Eye of the Poet: Studies in the Reciprocity of the Visual and Literary Arts from the Renaissance to the Present*. Academia.edu, 2002.

J. Gonzalez, S. Ashraf, A. Gelbukh, and P. Rosso. On the effect of individual annotator bias for subjective tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 636–644. European Language Resources Association, 2020.

M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

A. Hoyle, P. Goel, D. Peskov, et al. Is automated topic model evaluation broken?: The incoherence of coherence. *arXiv preprint arXiv:2107.02173*, 2021. URL `https://arxiv.org/abs/2107.02173`.

X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618, 2013.

J. D. Hunt. Ut pictura poesis, ut pictura hortus, and the picturesque. *Word & Image*, 1(1):87–107, 1985. doi: 10.1080/02666286.1985.10435668.

K. P. Johnson, P. Burns, J. Stewart, and T. Cook. Cltk: The classical language toolkit, 2014–2021. URL `https://github.com/cltk/cltk`.

R. B. Johnson and A. Creech. Likert-type scales: Using and interpreting likert-type scales. *Journal of Counseling and Development*, 73(3):384–386, 1995.

S. Kaleem, Z. Jalil, M. Nasir, and M. Alazab. Word embedding empowered topic recognition in news articles. *PeerJ Computer Science*, 10:e2300, 2024.

A. Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

S. Karst. *Schilderen in een land zonder schilders. De Nederlandse bijdrage aan de opkomst van de Britse schilderschool, 1520-1720.* PhD thesis, Utrecht University, 2021.

P. Kherwa and P. Bansal. Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, "7"(24), July 2019. ISSN 2032-9407. URL `https://eudl.eu/doi/10.4108/eai.13-7-2018.159623`.

L. Klein and J. Eisenstein. Reading thomas jefferson with topicviz: towards a thematic method for exploring large cultural archives. *Scholarly and Research Communication*, 4(3), 2013.

A. C. Kozlowski, M. Taddy, and J. A. Evans. The geometry of culture: Analyzing meaning through word embeddings. *CoRR*, abs/1803.09288, 2018. URL `http://arxiv.org/abs/1803.09288`.

L. Krušić. Constructing a sentiment-annotated corpus of austrian historical newspapers: Challenges, tools, and annotator experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62. Association for Computational Linguistics, 2024.

S. Kuzi, A. Shtok, and O. Kurland. Query Expansion Using Word Embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1929–1932, Indianapolis Indiana USA, Oct. 2016. ACM. ISBN 978-1-4503-4073-1. doi: 10.1145/2983323.2983876. URL `https://dl.acm.org/doi/10.1145/2983323.2983876`.

D. Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

F. J. Levy. Henry peacham and the art of drawing. *Journal of the Warburg and Courtauld Institutes*, 37:174–190, 1974. ISSN 00754390. URL http://www.jstor.org/stable/750839.

B. Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

H. Liu. Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*, 2017.

T. Liu, N. L. Zhang, and P. Chen. Hierarchical latent tree analysis for topic detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 256–272. Springer, 2014.

Z. Liu. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 2024. doi: doi: 10.1515/jtc-2023-0019. URL https://doi.org/10.1515/jtc-2023-0019.

S. Loria. Textblob: Simplified text processing. https://github.com/sloria/TextBlob, 2014.

G. F. Lytle and S. Orgel, editors. *Patronage in the Renaissance*. Princeton University Press, 1981. URL http://www.jstor.org/stable/j.ctt7zv2qb.

E. Manjavacas Arévalo and L. Fonteyn. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pages 23–36. Association for Computational Linguistics, Dec. 2021. URL https://aclanthology.org/2021.nlp4dh-1.4.pdf.

J. Marjanen, E. Zosa, S. Hengchen, L. Pivovarova, and M. Tolonen. Topic modelling discourse dynamics in historical newspapers. *arXiv preprint arXiv:2011.10428*, 2020.

I. McCulloh, J. Burck, J. Behling, M. Burks, and J. Parker. Leadership of data annotation teams. In *2018 International Workshop on Social Sensing (SocialSens)*, pages 26–31, 2018. doi: 10.1109/SocialSens.2018.00018.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889, 2009.

S. M. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Woodhead Publishing, 2016.

B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, 56(6):5133–5260, June 2023. ISSN 1573-7462. doi: 10.1007/s10462-022-10254-w. URL https://doi.org/10.1007/s10462-022-10254-w.

A. Myka and U. Güntzer. Fuzzy full-text searches in ocr databases. In N. R. Adam, B. K. Bhargava, M. Halem, and Y. Yesha, editors, *Digital Libraries Research and Technology Advances*, pages 131–145, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg. ISBN 978-3-540-68527-2.

L. K. Nelson. Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1):3–42, 2020.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

D. J. Newman and S. Block. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767, 2006.

F. Nugraha et al. Sentiment analysis techniques: A systematic literature review. *Tuijin Jishu/Journal of Propulsion Technology*, 2022.

X. Ouyang, P. Zhou, C. H. Li, and L. Liu. Sentiment analysis using convolutional neural network. In *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*, pages 2359–2364. IEEE, 2015.

Oxford English Dictionary. Oed thesaurus, 1857. URL `https://www.oed.com/thesaurus/start`. Accessed: January 30, 2024.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

R. Y. Pang and K. Gimpel. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. *arXiv preprint arXiv:1810.11878*, 2018.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

A. Poncelas, M. Aboomar, J. Buts, J. Hadley, and A. Way. A tool for facilitating OCR postediting in historical documents. In R. Sprugnoli and M. Passarotti, editors, *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 47–51, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-53-5. URL `https://aclanthology.org/2020.lt4hala-1.7/`.

ProQuest. Early english books online (eebo), 2015. URL `https://eebo.chadwyck.com`.

T. K. Rabb and J. Brown. The evidence of art: Images and meaning in history. *The Journal of Interdisciplinary History*, 17(1):1–6, 1986. ISSN 00221953, 15309169. URL `http://www.jstor.org/stable/204122`.

A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever. Better language models and their implications. *OpenAI blog*, 1(2), 2019.

H. Roose, W. Roose, and S. Daenekindt. Trends in contemporary art discourse: Using topic models to analyze 25 years of professional art criticism. *Cultural Sociology*, 12(3):303–324, 2018. doi: 10.1177/1749975518764861. URL https://doi.org/10.1177/1749975518764861.

F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*, 2014.

F. Sebastiani and A. Esuli. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th international conference on language resources and evaluation*, pages 417–422. European Language Resources Association (ELRA) Genoa, Italy, 2006.

S. Sia, A. Dalmia, and S. J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*, 2020.

J. Singh and P. Tripathi. Sentiment analysis of twitter data by making use of svm, random forest and decision tree algorithm. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 193–198. IEEE, 2021.

N. Singh and U. C. Jaiswal. A detailed sentiment analysis survey based on machine learning techniques. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 12(1):e29105–e29105, 2023.

R. Sprugnoli, S. Tonelli, A. Marchetti, and G. Moretti. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4):762–772, 2016.

A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

Y. W. Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

Text Creation Partnership. Early English Books Online Text Creation Partnership (EEBO-TCP). Available at: https://quod.lib.umich.edu/e/eebogroup/, 2015a. Accessed: January 30, 2024.

Text Creation Partnership. Eighteenth Century Collections Online Text Creation Partnership (ECCO-TCP). Available at: https://quod.lib.umich.edu/e/ecco/, 2019b. Accessed: February 27, 2024.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

D. Townsend. The picturesque. *The Journal of Aesthetics and Art Criticism*, 55(4):365–376, 1997. ISSN 00218529, 15406245. URL http://www.jstor.org/stable/430924.

K. Vukatana. Ocr and levenshtein distance as a measure of image quality accuracy for identification documents. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–4. IEEE, 2022.

M. Wankhade, A. C. S. Rao, and C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

A. Yadav and D. K. Vishwakarma. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 56:5731–5801, 2023.

P. Zhang, S. Wang, D. Li, X. Li, and Z. Xu. Combine topic modeling with semantic embedding: Embedding enhanced topic model. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2322–2335, 2019.

W. Zhao, J. Chen, X. Wu, et al. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 56:10661, 2023. doi: 10.1007/s10462-023-10661-7. URL `https://link.springer.com/article/10.1007/s10462-023-10661-7`.

P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016.

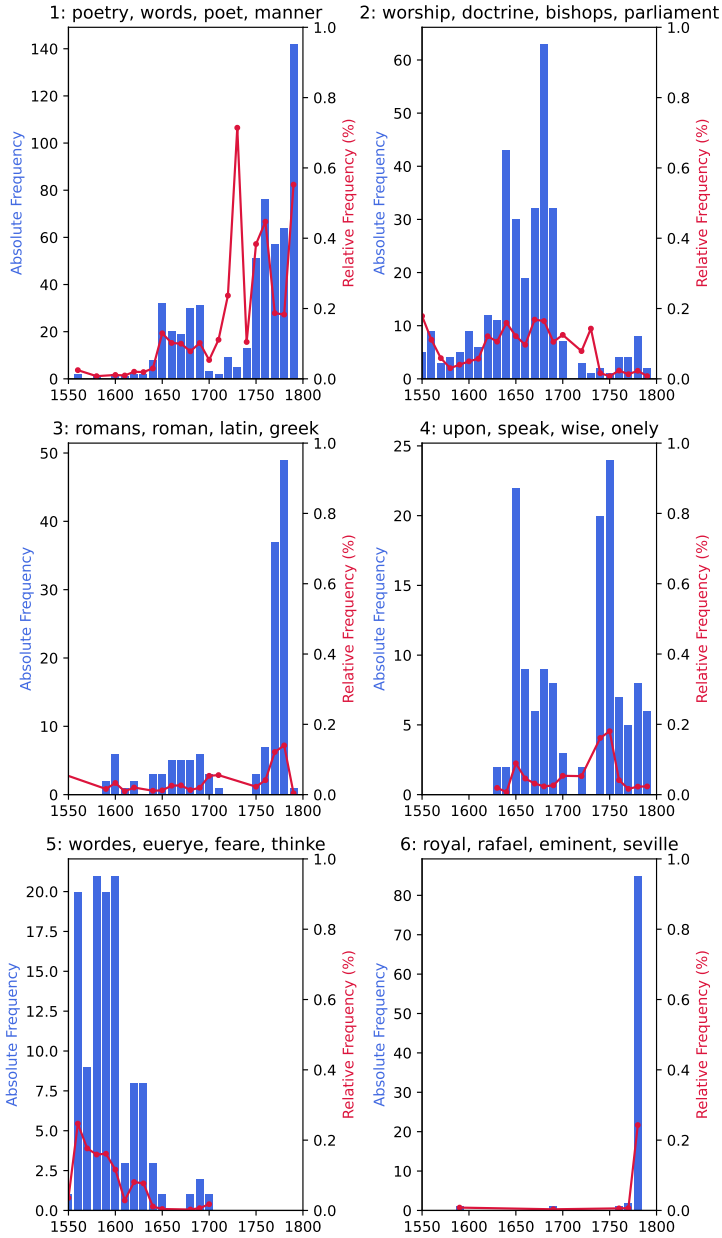
Q. Zhu, Z. Feng, and X. Li. Neural topic modeling with bidirectional adversarial training. *arXiv preprint arXiv:2004.12331*, 2020.

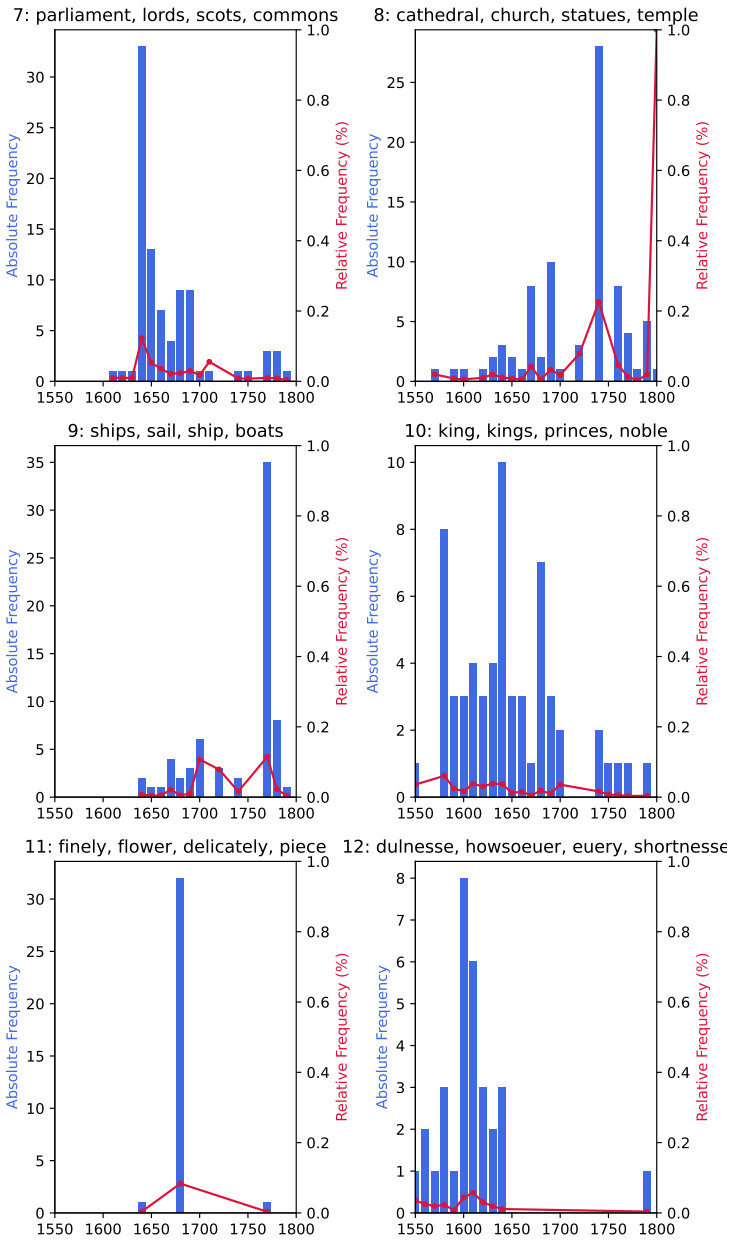
Appendix A

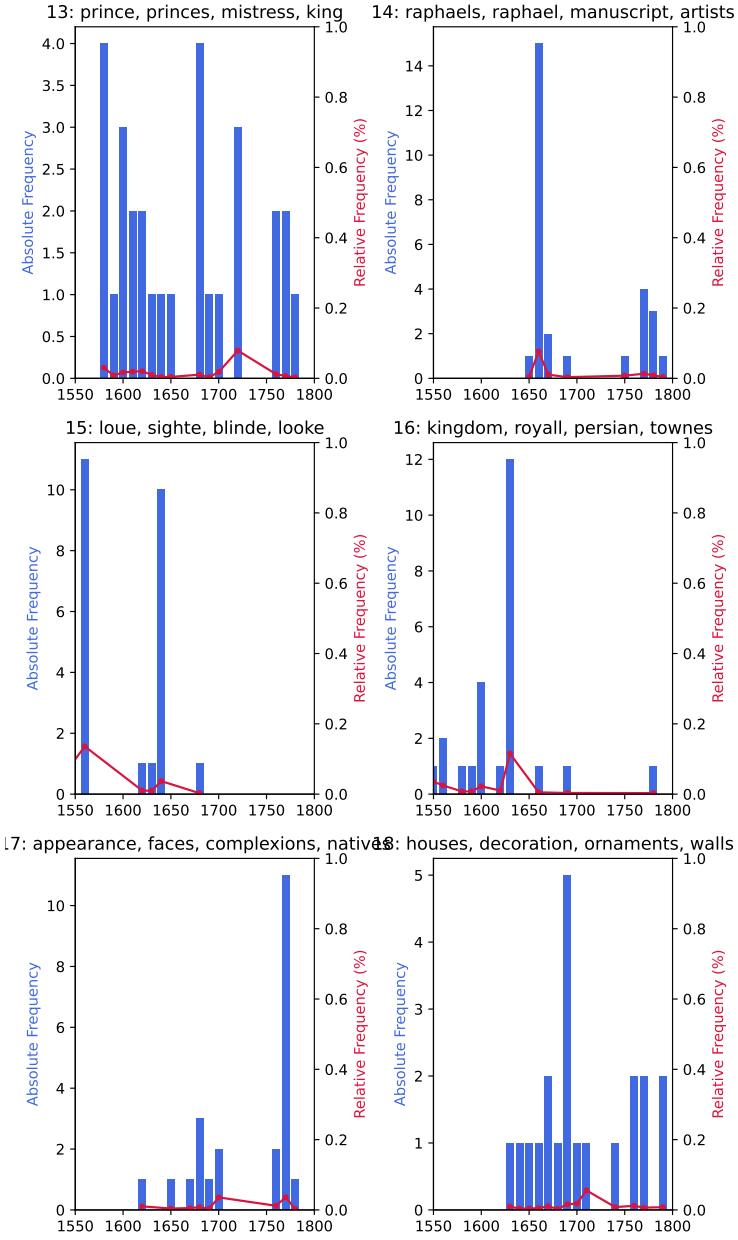
Complete plots of trend analysis

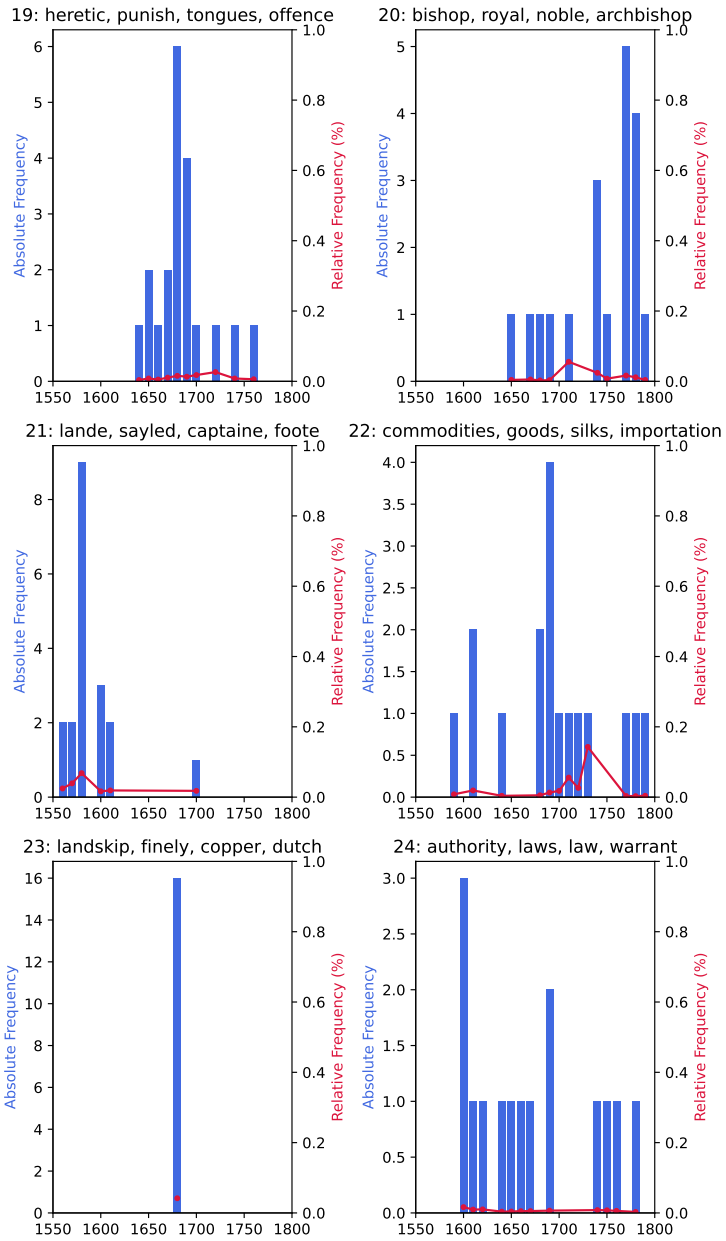
For the trend analysis, plots of the frequency and sentiment were made. In Figure A.1 the frequency plots for all topics are shown. The blue bars represent the absolute frequency per decade, which uses the y-ax on the left. The red represents the relative frequency per decade, the percentage of paragraphs in that topic for that decade, and uses the y-ax on the right. The same goes for Figure A.2, which represents the averaged sentiment per decade.

Figure A.1: Analysis of topic frequencies across decades. Blue represents the total frequency for each decade, and red indicates the percentage of paragraphs dedicated to each topic within that decade.




7: parliament, lords, scots, commons
8: cathedral, church, statues, temple
9: ships, sail, ship, boats
10: king, kings, princes, noble
11: finely, flower, delicately, piece
12: dulnesse, howsoeuer, euery, shortnesse
Absolute Frequency
Relative Frequency (%)
1550 1600 1650 1700 1750 1800


13: prince, princes, mistress, king
14: raphaels, raphael, manuscript, artists
15: loue, sighte, blinde, looke
16: kingdom, royall, persian, townes
17: appearance, faces, complexions, natives
18: houses, decoration, ornaments, walls
Absolute Frequency
Relative Frequency (%)


19: heretic, punish, tongues, offence
20: bishop, royal, noble, archbishop
21: lande, sayled, captaine, foote
22: commodities, goods, silks, importation
23: landskip, finely, copper, dutch
24: authority, laws, law, warrant
Absolute Frequency
Relative Frequency (%)
1550 1600 1650 1700 1750 1800

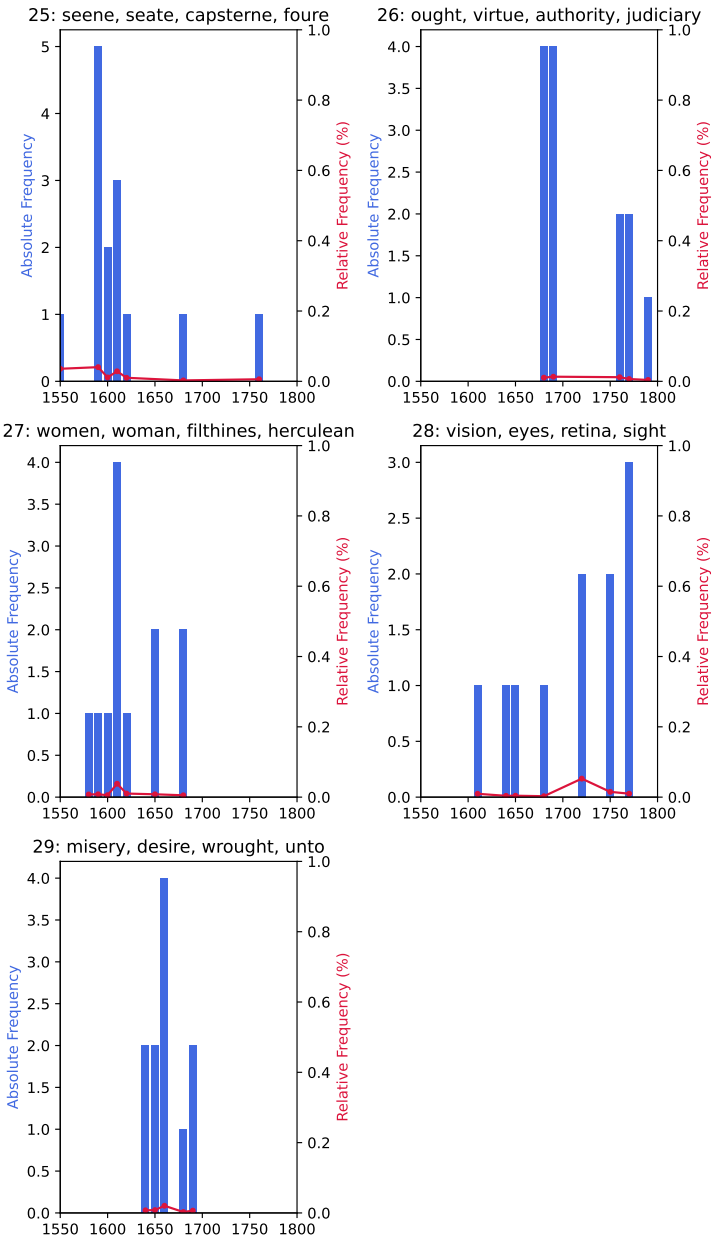
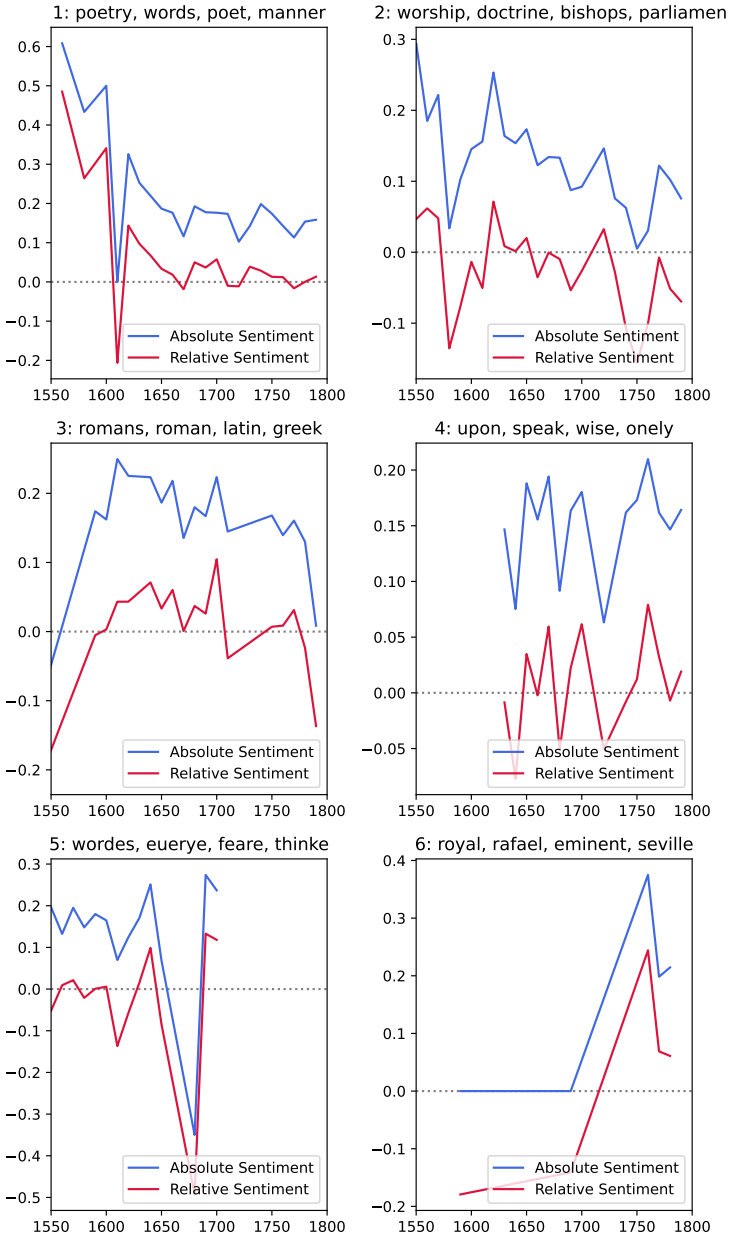
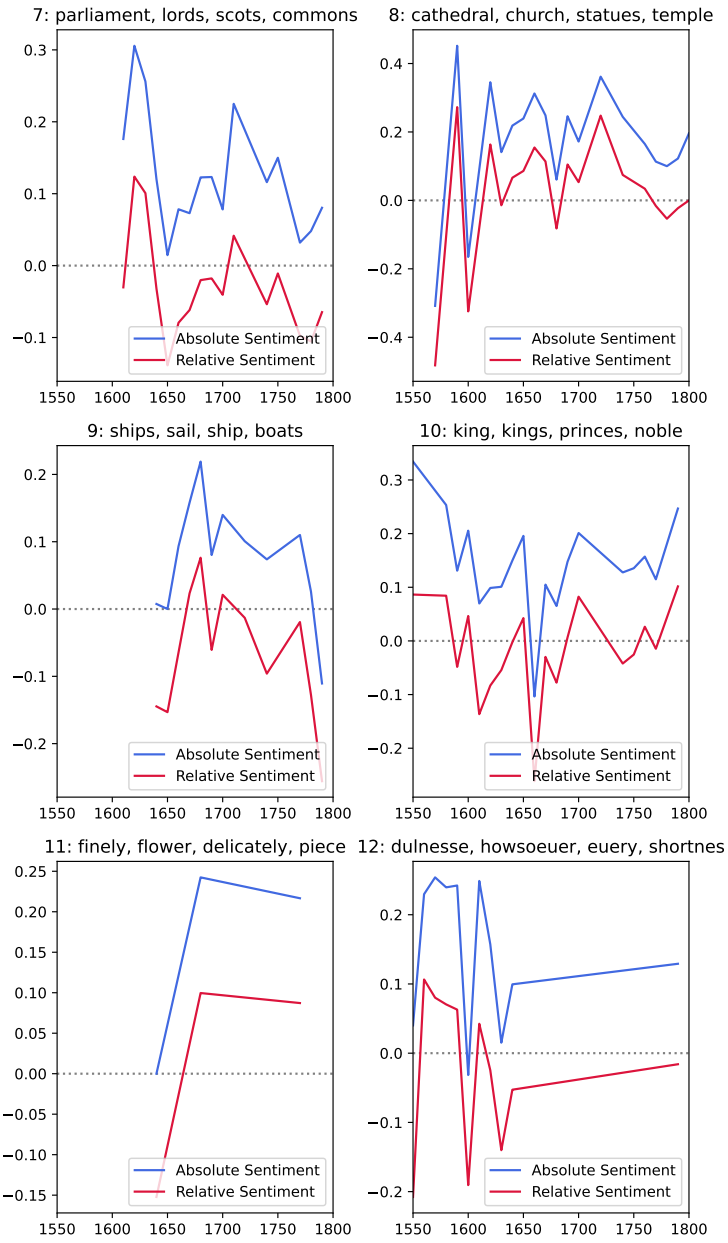
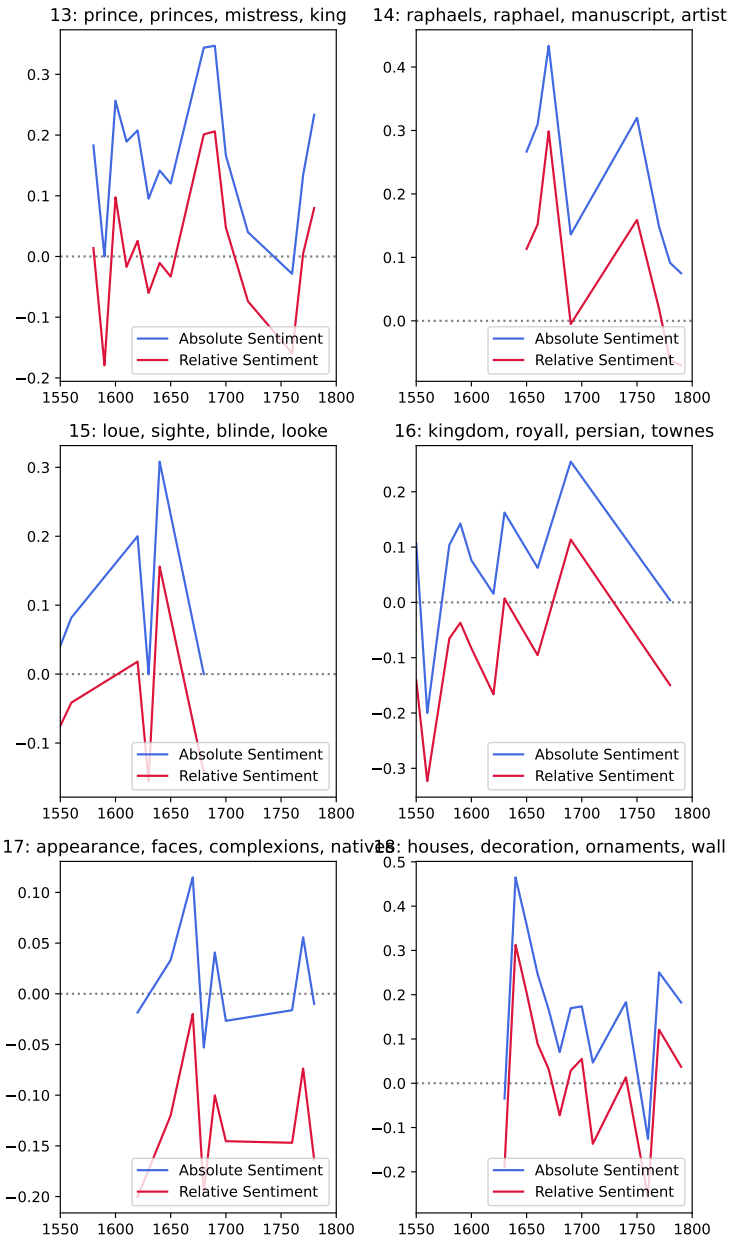
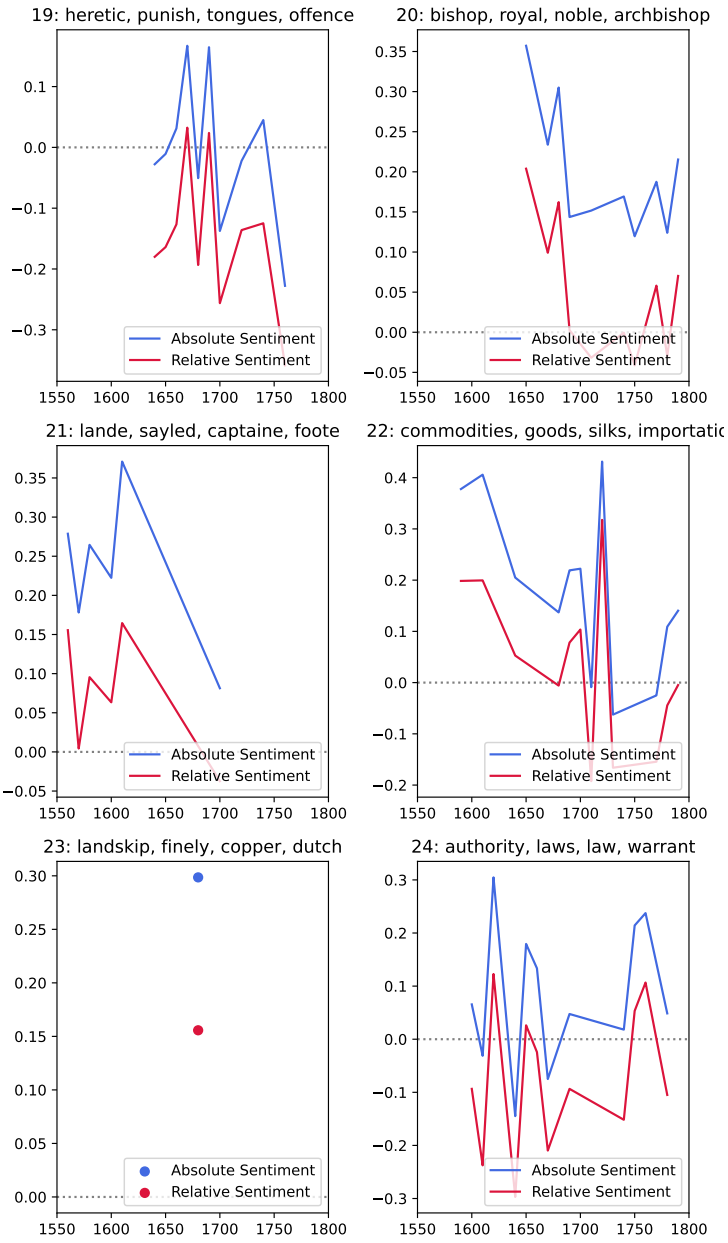

25: seene, seate, capsterne, foure
26: ought, virtue, authority, judiciary
27: women, woman, filthines, herculean
28: vision, eyes, retina, sight
29: misery, desire, wrought, unto
Absolute Frequency
Relative Frequency (%)
1550 1600 1650 1700 1750 1800

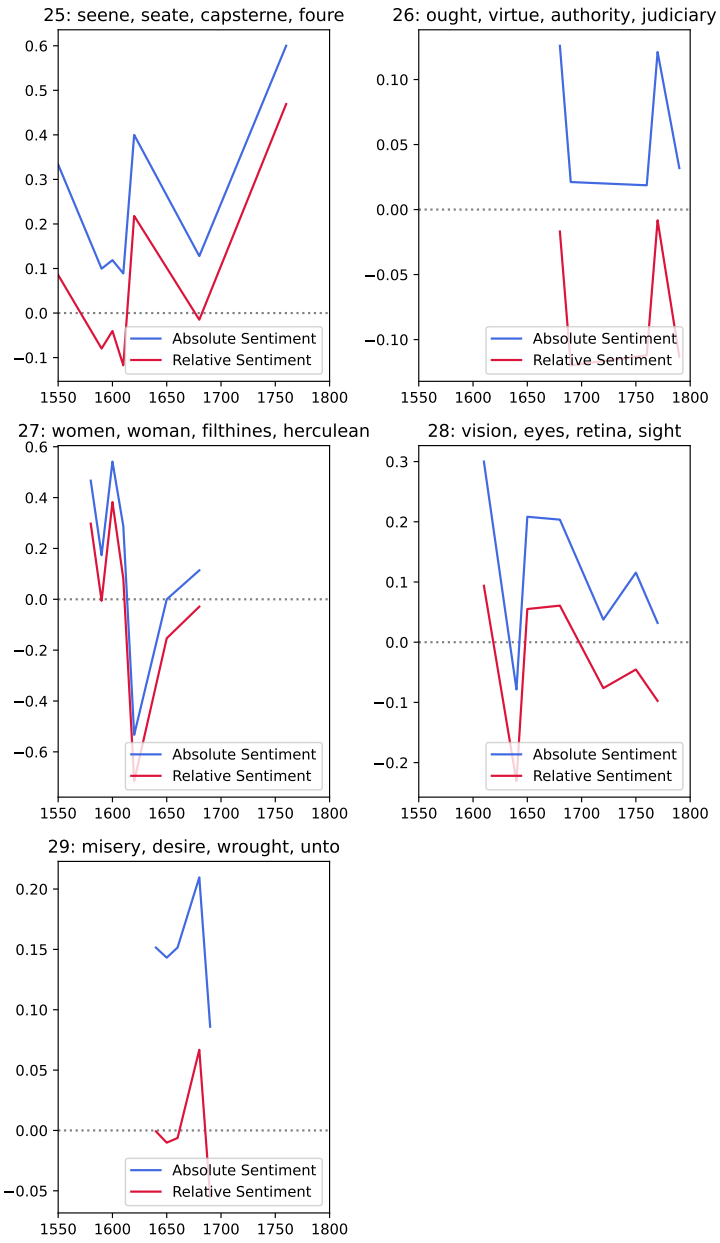
Figure A.2: The sentiment of different topics over the decades. Blue represents the average sentiment of each decade, and red indicates the average sentiment adjusted by subtracting the overall average sentiment of all paragraphs from that decade.




7: parliament, lords, scots, commons
8: cathedral, church, statues, temple
9: ships, sail, ship, boats
10: king, kings, princes, noble
11: finely, flower, delicately, piece
12: dulnesse, howsoeuer, euery, shortnes
Absolute Sentiment
Relative Sentiment


13: prince, princes, mistress, king
14: raphaels, raphael, manuscript, artist
15: loue, sighte, blinde, looke
16: kingdom, royall, persian, townes
17: appearance, faces, complexions, natives
18: houses, decoration, ornaments, wall
Absolute Sentiment
Relative Sentiment


19: heretic, punish, tongues, offence
20: bishop, royal, noble, archbishop
21: lande, sayled, captaine, foote
22: commodities, goods, silks, importatic
23: landskip, finely, copper, dutch
24: authority, laws, law, warrant
Absolute Sentiment
Relative Sentiment


25: seene, seate, capsterne, foure
26: ought, virtue, authority, judiciary
27: women, woman, filthines, herculean
28: vision, eyes, retina, sight
29: misery, desire, wrought, unto
Absolute Sentiment
Relative Sentiment
1550
1600
1650
1700
1750
1800

Appendix B

Development Set

In Table B.1 the whole development set used for surveys and examples is showed. The ID's are generated with the code available in GitHub. The topic and sentiment are assigned by the models from the framework.

Table B.1: The total development set with ID, year, text, assigned topic and assigned sentiment.

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02790	1762	Kent's method of embellishing a field, is admirable. It is painting a field with beautiful objects, natural and artificial, disposed like colours upon a canvas. It requires indeed more genius to paint in the gardening way. In forming a landscape upon a canvas, no more is required but to adjust the figures to each other: an artist who lays out ground in Kent's manner, has an additional task, which is to adjust his figures to the several varieties of the field.	poetry, words, poet, manner	0.07

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_03267	1790	It is not in ornaments that the merit of an argument consists. In viewing a portrait, the figure and features are the principal objects of contemplation; nor can the most elegant drapery, or beautiful colouring please the minds eye, where the artist deviates from nature and truth. The intent will excuse in your minds, the blemishes of the performance. The opinions, if not expressed, are conceived in the ardor of patriotism; in disinterested love for a community, where the writer received that life, which he thinks can only be well spent in its service.	poetry, words, poet, manner	0.31
text_01938	1674	But some will say after this, what Licence is left for Poets, certainly the same that good Poets ever tooke, without being faulty (for surely the best were so somtimes, because they were but men) and that Licence is Fiction, which kind of Poetry is like that of Landschap painting and poems of this nature, though they be not Vera ought to be Verisimilia.	poetry, words, poet, manner	0.47
text_03373	1796	Dryden traces the whole history of genius in a couplet, "What in nature's dawn the child admired, The youth endeavoured and the man ACQUIRED." Yet is it not always necessary that this admiration should be felt in childhood, or in youth, since accidental causes have frequently directed the pursuits of genius.Some instances are collected in Curiosities of Literature. Fourth edition, vol. 1. p. Caresses and coercion also, have made many a youth, a bright genius; patronage and poverty have stimulated men to become illustrious artists.	poetry, words, poet, manner	0.25

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02994	1765	Alas! Alas! It is but a too melancholy truth of the frailty of human nature, and a too visible proof of the decline of genius, that Hogarth, a painter, in his peculiar walk, rivalled by none, superior to all, should, to indulge a personal resentment, publish a foolish, trifling, insignificant print, to prove—what?—why, that he hated Mr. Churchill, and that his own abilities were quite decayed.	unto, therefore, selfe, nature	-0.10
text_02091	1682	Thus have I with a Pen, not Pencil drawn the Authors-Picture; and there is in some men's Styles as in Faces and Features, such peculiar Idioties and distinguishing Ayres from all others, that it is needless to write the Authors-name (as was, over dull painting, accustomed of old)—This is Cock, This a Bull.	unto, therefore, selfe, nature	-0.26
text_01103	1625	36. Michael Angelo, the famous Painter, painting in the Popes Chappell, the Portraiture of Hell, and damned Soules, made one of the damned Soules so like a Cardinall, that was his enemie, as euerie bodie, at first sight, knew it: Whereupon the Cardinall complained to Pope Clement, desiring it might be defaced; who said to him; Why, you know verie well, I haue power to deliuer a soule out of Purgatorie, but not out of Hell.	unto, therefore, selfe, nature	0.15
text_03966	1799	Once, on a summer's evening, I had the pleasure and delight of seeing the setting sun take his leave, and couch behind the Malvern hills, which are at an immense distance, yet are seen from Dowdswell; the rays of light, from the sun upon the clouds, threw such a glow and solemnity upon the earth, tinting it with so beautiful a variety, that, had there been but water to have finished the scene, a judicious painter might have formed a picture equal to the best of Claude Lauraine:	unto, therefore, selfe, nature	0.26

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02997	1765	Our poet had also some disputes with D****n L***h, his quondam printer; but, as that affair is so recent in every one's memory, and so generally known, I shall not give an account of it here. I must, however, beg leave to observe, that, however faultless our poet might be, in his quarrel with the painter, I can scarcely think he was entirely so in that with the printer; as Mr. L***h is a very honest, deserving man, and a very intelligent and good artist.	poetry, words, poet, manner	0.52
text_02739	1773	In the second line, vig'rous is marked as a property of the cedar: indeed all epithets, whether they precede or follow, require emphasis.—Pierce is noted as painting a quickness and boldness of vegetation, while the imagination is raised to a more than ordinary height, by particularising skies.	poetry, words, poet, manner	0.09
text_03788	1774	The subject of this piece is a VISION, containing a contest for superiority between Our lady Dame LIFE, and the ugly fiend Dame DEATH: who with their several attributes and concomitants are personified in a beautiful vein of allegorical painting. Dame LIFE is thus forcibly described.	unto, therefore, selfe, nature	0.10
text_03744	1771	The library is large. The most remarkable things are, John Trevisa's translation of Higden's Polychronicon, in 1387; the manuscript excellently wrote, and the language very good, for that time. A very neat Dutch missal, with elegant paintings on the margin. Another of the angels appearing to the shepherds, wish one of the men playing on the bagpipes. A manuscript catalogue of the old treasury of the college.	romans, roman, latin, greek	0.52

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02941	1789	In the palaces here, there are generally several indifferent pictures mixed with a few good ones—There are two Apostles out of four, painted by Carlo Dolci, in the Palazzo Riccardi, which I think invaluable; there is a Muse by the same in the Palazzo Corsini—	cathedral, church, statues, temple	0.10
text_03419	1795	But it is also necessary to acknowledge, that men of genius are often unjustly reproached with foibles. The sports of a vacant mind, are misunderstood as follies. The simplicity of truth may appear vanity, and the consciousness of superiority, envy. Nothing is more usual than our surprise at some great writer or artist contemning the labours of another, whom the public cherish with equal approbation. We place it to the account of his envy, but perhaps this opinion is erroneous, and claims a concise investigation.	poetry, words, poet, manner	0.08
text_01677	1585	Esope who painted to vs by Byrdes, Fishes, Serpents, foure footed beasts, the forme of an honest and safe lyfe, being taken with the enimie, and made subiect (with fooles) to misery, was wt other bondme offered to be sold to Xantus the Phylosopher, and being demaunded by Xantus what he could do, to that end he might therafter rate his price, changed not with his fortune his opinion: but answered, as if he were rather to be estéemed of the byer, he coulde do nothing.	unto, therefore, selfe, nature	0.33
text_00435	1662	The first triumphall Arch through which the King passed was erected in Leaden Hall street neer the end of Lime-street, which represented a Woman figuring Rebellion, with her attendant Confusion, in monstrous and deformed shapes. Opposite to her was a representation of Britains Monarchy with a prospect painting of his Majesties landing at Dover above it	unto, therefore, selfe, nature	0.05

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_03412	1795	A man of genius may, however, be rendered the most agreeable companion. Few artists but are eloquent on the art in which they excel. He is an exquisite instrument if the hand of the performer knows to call forth the rich confluence of his sounds. If,The flying fingers touch into a voice. D'Avenant.	poetry, words, poet, manner	0.42
text_01845	1652	The Satyrs of Varro, who was the Painter of the Life and of the Minde, would also afford us very grateful knowledges: For though most serious Philosophy were in those Satyrs, yet was it as it were on flowers, and as in a place for debauch, all painted and perfumed with the gallantry of those times.	unto, therefore, selfe, nature	0.12
text_02992	1765	His writings in favour of that administration that brought glory and honour to Great-Britain, and strengthened its interest; among others, had given offence to Hogarth, the ingenious, and truly comic painter, whose works will immortalise his name; who, having a place at court, as serjeant-painter, espoused the cause of that administration that brought inf–y and dish—-r to Britain, and that MADE THE PEACE OF 1763.	unto, therefore, selfe, nature	0.50
text_02825	1764	Again, you say, "Had Homer's Work been legislative, his Business would have been to deliver a more perfect and improved System in each Kind P. 29.." How do you know that? Has Homer himself told you so? Upon what Authority do you make Homer wiser than he was, and wiser than the Times had made him? "He painted what he saw and believed (says Dr. B.) and painted truly: the Fault lay in the Opinions and Manners of the Times: In the Defects of an early and barbarous Legislation, which had but half-civilized Mankind Dissert. p. 82.."	unto, therefore, selfe, nature	0.44

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_00248	1677	The private Houses of Maestreicht are generally covered with a black Slat, or Ardoise, otherwise not very beautiful. The Town house is fair, seated in one of the Piazza's, built of white Stone; it hath Nine large Windowes in a row on each side, and within is very well painted by Theodorus van der Schuer, who was Painter to the Queen of Sweden. In another Piazza is a Fountain, rows of Trees, and the great Church. This Town was besieged and taken from the King of Spain by the Confederate States, in the year 1632.	unto, therefore, selfe, nature	0.16
text_01009	1688	To intimate the Majesty, appearing in him from his very Birth, there was painted a young Lion, newly born, and having his Eyes sparkling, with these Words, Nascendo perspicax.	poetry, words, poet, manner	0.16
text_03521	1782	There are several paintings by de Vargas in the famous cathedral of Seville, particularly in the tower, which was his last work. Luis de Vargas was not less remarkable for his devotion, than for his talents, and, following the example of the great emperor Charles, he used at his private hours to deposit himself in a coffin, which he kept in his closet, and in that posture pursue his meditation upon death: This event, for which he used such edifying preparation, took place in the year 1590.	royal, rafael, eminent, seville	0.17

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02989	1765	A pitiful ambition of displaying superior knowledge in particular arts and sciences, that have been canvassed in company, is very common, and is very ridiculous. This fault Churchill was never guilty of; and, tho' his erudition and knowledge were superior to most people's, and might have justified him in enforcing his sentiments with a proper warmth, and to take up a longer time than persons of inferior abilities could arrogate, he ever delivered his opinion in a decent manner, and was as patient an auditor, as he was a skilful orator.	unto, therefore, selfe, nature	0.29
text_02418	1662	Henry Golzius was a Hollander, and wanted only a good, and judicious choice to have render'd him comparable to the profoundest Masters that ever handled the Burin, for never did any exceed this rare workman; witnesse, those things of his after Gasparo Celio, the Gallatea of Raphael Santio, and divers other pieces after Polydor da Carravaggio, a Hierom, Nativity, and what he did of the Acts of the Apostles, with Ph. Galle, &c. but he was likewise an excellent painter.	raphaels, raphael, manuscript, artists	0.38
text_03666	1790	Andrew Brice, in figure and tremulous manner, was exactly what Mr. King appears to be in Lord Ogleby; and I could have forgiven Brice had he painted like his Lordship: for he had so much of the lily in his complexion, that he looked (tho' one of the neatest) the most corpse-like Mandarine figure I ever beheld in the various productions of Human Nature.	unto, therefore, selfe, nature	0.19
text_01010	1688	To signify, that the Birth of his Royal Highness is an Obstacle to the Designs of such, as would oppose Great Britains Happiness, was painted a rising Sun, with many Stars about it, whose Courses contrary to the Suns, are checke by his Arisal: The Motto was, Cursusque vagos statione moratur.	unto, therefore, selfe, nature	0.50

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_00246	1677	This City was made an Episcopal See, 1559. The Cathedral is Dedicated to St. John. In the Quire are painted the Arms of many of the Knights of the Golden Fleece. And over the upper Stalls or Seats, an Inscription in French, which contains the History of the first Institution, and Model of this Order, by the most High and mighty Prince Philip the Good, Duke of Burgundy, Lorain, and Brabant: Besides divers Statua's and Pillars. There are also several Monuments of the Bishops of Bosche and others.	cathedral, church, statues, temple	0.28
text_03434	1795	I would ask why the art of writing is not deserving of the same regard as the art of painting? And then I would enquire, what painting can urge in it's own cause, which will entitle it to a superiority over the art of composition?	poetry, words, poet, manner	0.10
text_01565	1634	As for the Natiues, they are proper tall men of person; swarthy by nature, but much more by Art: painting themselues with colours in oyle, like a darke Red, which they doe to keepe the Gnatts off: wherin I confesse, there is more ease then comlinesse.	unto, therefore, selfe, nature	0.25
text_02548	1728	THIS was the only Ship the English lost in this long Engagement. For although the Katherine was taken, and her Commander, Sir John Chicheley, made Prisoner, her Sailors soon after finding the Opportunity they had watch'd for, seiz'd all the Dutch Sailors, who had been put in upon them, and brought the Ship back to our own Fleet, together with all the Dutch Men Prisoners; for which, as they deserv'd, they were well rewarded. This is the same Ship which the Earl of Mulgrave (afterwards Duke of Buckingham) commanded the next Sea Fight, and has caus'd to be painted in his House in St. James's Park.	ships, sail, ship, boats	0.08

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_03011	1795	Those regions of decorated beauty being now forbidden ground, we confined our walks to some pasturage lands near the town, which were interspersed with a few scattered hamlets, and skirted by hills, and were so unfrequented, that we heard no sounds except the sheep-bell, and the nightingales, and saw no human figure but an old peasant with a white beard, who together with a large black dog took care of the flock. It was in these walks that the soul, which the scenes of Paris petrified with terror, melted at the view of the soothing landscape, and that the eye was lifted up to heaven with tears of resignation mingled with hope. I have no words to paint the strong feeling of reluctance with which I always returned from our walks to Paris, that den of carnage, that slaughterhouse of man. How I envied the peasant his lonely hut! for I had now almost lost the idea of social happiness. My disturbed imagination divided the communities of men but into two classes, the oppressor and the oppressed; and peace seemed only to exist with solitude.	unto, therefore, selfe, nature	0.09
text_02605	1783	As to the Scene, it is clear, that it must always be laid in the country, and much of the Poet's merit depends on describing it beautifully. Virgil is, in this respect, excelled by Theocritus, whose descriptions of natural beauties are richer, and more picturesque than those of the otherWhat rural scenery, for instance, can be painted in more lively colours, than the following description exhibits?	unto, therefore, selfe, nature	0.18

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02612	1783	IT deserves attention too, that in describing inanimate natural objects, the Poet, in order to enliven his description, ought always to mix living beings with them. The scenes of dead and still life are apt to pall upon us, if the Poet do not suggest sentiments, and introduce life and action into his description. This is well known to every Painter who is a master in his art. Seldom has any beautiful landscape been drawn, without some human being represented on the canvas, as beholding it, or on some account concerned in it: Hîc gelidi fontes, hîc mollia prata, Lycori, Hîc nemus; hîc ipso tecum consumerer aevoHere cooling fountains roll thro' flow'ry meads, Here woods, Lycoris, lift their verdant heads, Here could I wear my careless life away, And in thy arms insensibly decay. VIRG. Ecl. X. WARTON..	poetry, words, poet, manner	0.02

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02999	1790	On the evening of their arrival at the family-seat, Julia walked out with Charlotte, and felt, with particular sensibility, the beauties of nature. She had, till now, only seen the rich cultivated landscapes of the south of England; but her ardent imagination had often wandered amidst the wild scenery of the north, and formed a high idea of pleasure in contemplating its solemn aspect; and she found that the sublime and awful graces of nature exceed even the dream of fancy. The setting sun painted the glowing horizon with the most refulgent colours: immediately above its broad orb, which was dazzling in brightness, hung a black cloud that formed a striking contrast to the luxuriant tints below: some of the hills were thrown into deep shadow, others reflected the setting beams. When the sun sunk below the horizon, every object gradually changed its hue. The form of the surrounding hills, and the shape of the darkening rocks that hung over the lake, became every moment more doubtful; till at length twilight spread over the whole landscape that pensive gloom so soothing to an enthusiastic fancy. Every other sound was lost in the fall of the torrent, a sound which Julia had never heard before, and which seemed to strike upon her soul, and call forth emotions congenial to its solemn cadence.	poetry, words, poet, manner	0.12

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02760	1762	To illustrate this difference, I give the following examples. It has been explained why smoke ascending in a calm day, suppose from a cottage in a wood, is an agreeable objectChap. 1.. Landscape-painters are fond of this object, and introduce it upon all occasions. As the ascent is natural and without effort, it is delightful in a calm state of mind. It makes an impression of the same sort with that of a gently-flowing river, but more agreeable, because ascent is more to our taste than descent. A firework or a jet d'eau rouses the mind more; because the beauty of force visibly exerted, is superadded to that of upward motion. To a man reclining indolently upon a bank of flowers, ascending smoke in a still morning is delightful. But a fire-work or a jet d'eau rouses him from this supine posture, and puts him in motion.	unto, therefore, selfe, nature	0.42
text_03448	1774	Claude in his happiest hour never struck out a finer landskip; it has every requisite which the pencil can demand, and is perhaps the only view in England which can vie with the sublime scenes from which that painter formed his taste.	unto, therefore, selfe, nature	0.00
text_03472	1787	In the same year, 1768, upon the establishment of the royal academy of painting, sculpture, c. Johnson was nominated professor of ancient literature, an office merely honorary, and conferred on him, as it is supposed, upon the recommendation of the president, Sir Joshua Reynolds.	unto, therefore, selfe, nature	-0.25
text_02474	1640	Q. Wherefore doe some paint Love with the face of a man, and not of an Infant?	loue, sighte, blinde, looke	0.50
text_00293	1689	95 a stone Head and Flowers rarely painted by Brughel	unto, therefore, selfe, nature	0.30

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_00299	1689	131 a neat Landskip finely painted on board	landskip, finely, copper, dutch	0.42
text_00311	1689	174 a Merry piece finely painted by Haemskirk	finely, flower, delicately, piece	0.42
text_01607	1566	Bicause when he had painted Helena, he saide that Leda her mother for all that she was gotten with childe by Iupiter, had not made Helena so fayre as he had painted her.	wordes, euerye, feare, thinke	0.00
text_00327	1689	241 an Italian Landskip very curiously painted	landskip, finely, copper, dutch	-0.07
text_00349	1689	324 a Landskip finely painted by a Dutch master	landskip, finely, copper, dutch	0.42
text_03449	1774	—When the long protracted shadows of the mountains cast on the bosom of the Lake, shewed the vastness of those masses from whence they proceeded; and still as the moon arose higher in the horizon, the distant objects began to be illumined, and the whole presented us with a noble moonlight piece, delicately touched by the hand of nature; and far surpassing those humble scenes which we had often viewed in the works of the Flemish painters.	unto, therefore, selfe, nature	0.06
text_01548	1622	A painted woman is no perfect woman, for all women by nature are either faire or foule: but if an artificial faire be set vpon a foule complexion, it makes but a speaking picture, and a picture is no perfect woman.	women, woman, filthines, herculean	-0.53

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02557	1762	The greater the luxuries of every country, the more closely, politically speaking, is that country united. Luxury is the child of society alone, the luxurious man stands in need of a thousand different artists to furnish out his happiness: it is more likely, therefore, that he should be a good citizen who is connected by motives of self-interest with so many, than the abstemious man who is united to none.	poetry, words, poet, manner	0.38
text_03536	1782	Juan del Castillo of Seville was a painter of eminence and in great repute as a master and instructor in the art; he had the double honour of being disciple of Luis de Vargas, and teacher of Bartolome Murillo; the famous Alonso Cano, and Pedro de Moya were likewise his scholars: He died at Cadiz, aged 56, in the year 1640.	royal, rafael, eminent, seville	0.30
text_02918	1797	Schedoni, on the contrary, advanced in years, exhibited a severe physiognomy, furrowed by thought, no less than by time, and darkened by the habitual indulgence of morose passions. He looked as if he had never smiled since the portrait was drawn; and it seemed as if the painter, prophetic of Schedoni's future disposition, had arrested and embodied that smile, to prove hereafter that cheerfulness had once played upon his features.	poetry, words, poet, manner	0.10
text_01197	1636	Nay therefore (quoth Dorastus) maids must loue because they are young: for Cupid is a child, and Venus, though old, is painted with fresh colours.	wordes, euerye, feare, thinke	0.17
text_00410	1697	Let us see if it be possible to prescribe some rule of Love, which is often what makes Marry'd People most Unhappy; sometimes because it is wanting, and sometimes because it is excessive. Let us at least spread the Nets to catch this sort of prudent Love, and let him fall into the Snare if he will, though it is likeliest he will fly from it, and that perhaps is the reason he is painted with Wings.	heretic, punish, tongues, offence	0.14

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_00525	1650	O imperious and impious Love, thou deluding Traytor, how rightly did the Poets and Painters, paint thee blind, and naked? Since thou hast no eyes to see into how many dangers thou leadest thy servants; and like thy self, makest them both blind and naked, disrobeing them of all their vertuous abiliments, that their naked shame may appear in their found pursuits.	poetry, words, poet, manner	0.04
text_02420	1662	Siiderhoef has engraven the heads of most of the Learned Dutch, after several painters with good successe: as those of Heinsius, Grotius, Barleus, c. not forgeting that stupendious Lady Anna Maria a Scureman, c.	raphaels, raphael, manuscript, artists	0.40
text_00088	1612	A faire morning commonly betokeneth a faire and a pleasant day; and the good companie, which a man frequenteth, is a strong argument that he is disposed well. It is for painters to deuise for their pictures such visages, and such faces, as they please, but you may not chuse whom you list, for your familiar consorts and companions.	unto, therefore, selfe, nature	0.28
text_01597	1566	I desire to knowe wherfore the notable painter Zeuxis did painte him with a grene robe?	unto, therefore, selfe, nature	0.50

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_02947	1789	I Got away as fast as I possibly could from Vienna; for if I had staid a week longer, I am convinced I should have staid the whole winter. The country between Vienna and Cracow is very fine; chiefly open, here and there the plain beautifully varied with hills of gentle ascent, and small woods; the sportsman and the painter would be pleased with it, as it affords a variety of landscapes and game, equally favourable to both. The firs and deciduous trees do not seem to flourish in the same spot; I frequently saw a wood of the one to my right, and of the other to my left; I observed, that cattle of all sorts are suffered to eat the green corn during the hard weather—	unto, therefore, selfe, nature	0.07
text_03435	1795	There are two opinions relative to the state of men of genius. One party imagine that no protection from the great, or a court, is necessary for the encouragement of artists; and the other are persuaded, that when honours and pensions are judiciously distributed, it excites emulation in the young, and gives that leisure to those on whom they are bestowed, so necessary to some, to cultivate their talents. They think with Boileau, that Un AUGUSTE aisement peut faire des VIRGILES.	dulnesse, howsoeuer, euery, shortnesse	0.13

Continued on next page

Table B.1 – continued from previous page

Text ID	Year	Text	Assigned Topic	Assigned Sentiment
text_00381	1660	Her Face would make a fit Meddall in Copper, upon which for a Motto there should be in Capitalls ingraved IMPUDENCE. Her eyes are two gogling sparkling Globes, between which and her mouth there's a pretty speckled painted jewel, from whence there daily flows a Cream-pot of Nectar. By her Omega's nose, she should have been Related to the late Tyrant Oliver; and its no great wonder to see such a copper Nose upon a brazen Face: or, by some other parts of her, she should have been related to the Hairy woman, I warrant one might get many a pound by shewing her naked at Market Towns and Fairs; for no man can easily imagine there's such a Monster in Nature, unless he see it.	unto, therefore, selfe, nature	0.06
text_02581	1783	NOW this high power which eloquence and poetry possess, of supplying Taste and Imagination with such a wide circle of pleasures, they derive altogether from their having a greater capacity of Imitation and Description than is possessed by any other art. Of all the means which human ingenuity has contrived for recalling the images of real objects, and awakening, by representation, similar emotions to those which are raised by the original, none is so full and extensive as that which is executed by words and writing. Through the assistance of this happy invention, there is nothing, either in the natural or moral world, but what can be represented and set before the mind, in colours very strong and lively. Hence it is usual among critical writers, to speak of Discourse as the chief of all the imitative or mimetic arts; they compare it with painting and with sculpture, and in many respects prefer it justly before them.	poetry, words, poet, manner	0.13