Comparing Feature Engineering Techniques for the Time Period Categorisation of Novels

Fereshta Westin

University of Borås, Allégatan 1, Borås, Sweden
fereshta.westin@hb.se

Fereshta Westin holds a Master's degree in Informatics and is currently pursuing a Ph.D. in Library and Information Science at the University of Borås, Sweden. Her research focuses on organizing novels based on their temporal information, such as the time period in which the story is set, including eras like the Medieval times or World War I. To achieve this, she employs machine learning and large language models for organizing temporal data. Her research interests lie at the intersection of knowledge organization and machine learning.

Westin, Fereshta. 2024. "Comparing Feature Engineering Techniques for the Time Period Categorisation of Novels". *Knowledge Organization* 51, no.5: 330-339. 25 references. DOI:10.5771/0943-7444-2024-5-330.

Abstract: The growing number of literary works being produced and published has emphasised the importance of better cataloguing methods to handle the increasing volume effectively. One specific issue is the lack of organising works by time periods, which is crucial for understanding and organising literature. In this study, "time" refers to when the story's events occur or the narrative's temporal setting, like specific historical periods or events, rather than the publication date. Categorising literary works based on their historical settings can significantly improve accessibility for library patrons navigating online catalogues. However, time period categorisation is uncommon, primarily due to the resource-intensive nature of the process, which necessitates extensive analysis by librarians and cataloguers. To address this issue, this paper proposes evaluating different machine learning workflows to predict time periods for novels. The workflow comprises preprocessing, feature engineering, classification, and evaluation. The feature engineering techniques used are Latent Dirichlet Allocation (LDA), Word Embedding with Sentence-BERT (WE SBERT), and Term Frequency-Inverse Document Frequency (TF-IDF), and the classification algorithm used is Logistic Regression. The models are assessed using the F1 score, precision, and recall metrics. The time period categories used are Medieval, Era of Great Power, Age of Liberty, and Gustavian periods. The objective is to determine how effectively each model categorises Swedish historical fiction novels into their appropriate time period categories. By leveraging machine learning techniques, the research seeks to supplement the time period categorisation process, aiding cataloguers and ultimately enhancing the accessibility and usability of library collections.

Received: 20 March 2024; Revised 04 April 2024; Accepted 25 April 2024.

Keywords: literary categorization; machine learning techniques; time period categorization.

† This article was selected as one of the best papers at the Seventh ISKO UK biennial Conference, July 24-25, 2023, held at the University of Strathclyde, Glasgow, in partnership with the School of Computing, Engineering & the Built Environment at Edinburgh Napier University. This is an extended version of the article published in the proceedings of the ISKO UK conference.

1.0 Introduction

As digitisation technology has become more widely used in library operations, Gartner (2008) states that the necessity for metadata has become increasingly pressing. Guerrini (2023) explains that metadata serves as both a representation of documents and an infrastructure for searching and traversing networked resources. It provides essential information about documents, including title, author, subject, and publication date, acting as a summary or description of the document's content and attributes. Simultaneously, metadata serves as a foundational framework within digital environments, such as digital libraries or databases, enabling searching and navigating networked resources. By organising information in a structured manner, metadata allows, for example, patrons at a library to locate specific documents or resources. While bibliographic elements such as title, author, and subject have long been the primary search criteria in online public access catalogues (OPAC), Boogard et al. (2019) claim that evolving user search behaviour now encompasses more general queries, including time periods and events, thus creating new challenges in accessing digital

library resources. Specific time periods or events are often sought after, posing unique difficulties in digital search queries (Petras et al., 2006). Furthermore, Frommeyer (2013) argues that time is an important component of subject cataloguing, and a study by Bates et al. (1993) found that 16 percent of humanities scholars' searches on the DIALOG database were related to time.

As Petras et al., (2006) points out, the conventional method of searching by dates in OPACs fails to capture the nuances of historical epochs effectively where users are often limited to conducting keyword date searches, such as entering "1700," which retrieves literature both written in and about the year 1700. Moreover, this keyword search only works if the searched term is specifically mentioned, thereby limiting its effectiveness in capturing comprehensive sets of literature related to a particular historical period. This method lacks specificity and may yield results unrelated to the user's intended historical period. An alternative and more nuanced approach to organising literature involves analysing the texts' content to derive the time periods they represent rather than relying solely on keyword searches based on exact matches of keywords. This method entails extracting temporal information directly from the narrative content, allowing for a comprehensive and contextually accurate categorisation of literature into specific historical periods.

In this study, the term "time" denotes the historical period during which the events of a narrative unfold. "Time period" and "historical period" are utilised interchangeably to articulate this concept; similarly, the terms "literary work" and "novel" are used interchangeably. This study emphasises contextualising the story within a broader historical framework rather than the publication date of a novel. Most fiction and historical novels have at least one time period where the story unfolds, and this temporal information can be retrieved and used as metadata. For example, suppose a novel is set during the Medieval period; it immerses readers into that era's cultural, social, and political milieu, offering insights into the customs, beliefs, and challenges of the time. Similarly, narratives set during the Viking Age transport readers to a time of exploration, conquest, and cultural exchange in Northern Europe.

The National Library of Sweden and its Metadata Office are responsible for maintaining the guidelines for categorising literary works by time periods; educated cataloguers at different libraries are responsible for this categorisation. In Sweden, the joint library catalogue (Libris), which cataloguers use, has around nine million printed works of fiction and two million printed works of non-fiction. I have crafted a fictional scenario featuring a fictional character to exemplify how a time period search can be executed in systems like Libris. This case aims to illustrate the process of conducting a time period search. Case: Alex is an avid reader fascinated by fictional stories set in the medieval period. He enjoys tales of castles, knights, armies, and brave individuals fighting for justice and freedom. Although Alex prefers stories grounded in historical facts, he seeks fictional narratives rather than factual accounts. To find such stories, Alex utilises an online system specifically designed for literature collections rather than a general search engine like Google. He begins his search by entering keywords such as "medieval period," hoping to uncover exciting tales from that era. However, he quickly becomes overwhelmed by the sheer number of results returned by the search. Despite attempting alternative keywords like "medieval fictional books," Alex fails to find satisfactory results. Adding to his frustration, many search results consist of factual books rather than the fictional narratives he desires. Regarding the browsing feature, Alex explores categories and subjects but struggles to find relevant content. The categories and subjects provided are too broad, making it difficult for Alex to navigate effectively and locate fictional books set in the medieval period. As Boogard et al. (2019) also suggest, metadata representing time periods can be added to search engines in online collections, such as library systems, to increase the chances of a user finding a novel set in a specific historical period. Searching and browsing for time periods, such as the medieval period, poses several challenges in online systems dedicated to literature collections. Unlike search engines like Google, which rely on algorithms to index and rank web pages, specialised online systems for literature collections often use metadata and keywords. Additionally, the categorisation and subject tagging in these systems may be too generalised, making it challenging for users like Alex to narrow their search to find specific fictional narratives. As a result, users may struggle to locate the desired content efficiently, leading to frustration and dissatisfaction with the search experience.

The scarcity of relevant findings in systems like Libris can be because only a fraction of its extensive collection includes time period metadata. A database search in Libris revealed that out of the 12 million works available, only 108,816 had specific time periods specified. These numbers support the claim made in Dalli's (2006) study that finding literary works with time period metadata is rare, making time period searches difficult. Although it is possible to categorise literary works by time periods, they are often not prioritised, especially in the case of fiction. A possible reason for the lack of time period metadata could be that it is time-consuming and challenging to decide what metadata to specify in each case. Cataloguers categorising in Libris can obtain information on a time period from the author, publishers, other cataloguers, or organisations. Without information on the time period, cataloguers must read the blurb or a portion of the text, search the internet, or ask colleagues. Hence, understanding the text characteristics is a crucial step, and it is where cataloguers must spend most of

their time when cataloguing. A poor understanding of the work's characteristics can lead to mis-categorisation. Therefore, a commonly used principle for creating metadata that represents the time periods of a literary work is that the work must contain a minimum of 20 percent of its content dedicated to describing one or more time periods (Metadatabyrån 2021b).

Feature engineering is a necessary step in machine learning, where raw data is transformed into meaningful features that can be used to train models and make predictions (Håkansson and Hartung 2020). This process involves transforming data attributes to improve the performance of machine learning algorithms. Likewise, when cataloguing literature, cataloguers comprehend each work's attributes and annotate them accordingly to identify patterns in data and determine their subject matter, themes, and historical context. By understanding these text characteristics, cataloguers can assign appropriate metadata such as genre, time period, and subject tags.

The effectiveness of both feature engineering in machine learning and cataloguing by librarians or cataloguers hinges on the quality of text analysis. In machine learning, prediction accuracy hinges on various factors, including the preprocessing methods applied to the texts, the types of techniques employed to generate features, and the selection of algorithms used for text categorisation. While in human cataloguing, the precision of categorisation relies on the cataloguer's ability to accurately interpret and classify literary works' content. Thus, both processes require a deep understanding of the data or text being analysed to achieve optimal results.

This study aims to compare different feature engineering techniques alongside preprocessing and classification. Then, the models are evaluated to determine the most accurate approach for predicting time periods in Swedish historical fiction literature sourced from the Swedish Literature Bank. Machine learning uses algorithms and statistical models to perform a specific task without being explicitly instructed; it relies on statistically finding patterns and inferences by looking at many sample data, also called training data. When the model has "learned" the patterns in the training data, it can start making predictions on data it has never seen. Integrating machine learning techniques into the cataloguing process holds the potential to enhance the precision and efficiency of time period categorisation by aiding cataloguers in their decision-making. The feature engineering techniques that are being compared are Latent Dirichlet Allocation (LDA), Word Embedding with Sentence-BERT (WE SBERT), and Term Frequency-Inverse Document Frequency (TF-IDF). Logistic regression is used to classify the Swedish historical fiction texts into one of these time periods: Medieval period, Era of Great Power, Age of Liberty and, Gustavian Period. The models are evaluated by F1 score, together with precision and recall. Further details are outlined in the Methods section.

2.0 Fiction categorisation

The topic of fiction content analysis and retrieval has become increasingly significant in the context of knowledge management and literature organisation, as pointed out by Saarti (2019). However, Saarti (2019) continues that categorising fiction has proven challenging due to its multifaceted and interpretive nature. Rafferty (2013) observes that genre has traditionally been employed to categorise fiction, owing to its usage in advertising and targeted marketing (Saarti 2019; Maker 2008). Shenton (2006) outlines a project that sought to establish new categories for fiction within a high school library based on an analysis of the book's nature in the collection. These categories were informed by the content of a six-month log of fiction inquiries. More recently, Almeida and Gnoli (2021) claim that conventional categorisation systems that index fictional works based only on their form, genre, and language may not be the most effective approach since they need to consider the story's actual content. Therefore, it is essential to develop new methods to better analyse and categorise fiction content, as traditional categorisation systems have been found to be inadequate in this regard. There are numerous ways of categorising fiction, as evidenced by various attempts. However, one prominent approach involves analysing the content. Machine learning can simplify this task by enhancing the ease of conducting content analysis. Manger (2018) explored the possibilities of using machine learning to categorise a text as fiction or nonfiction by analysing reviews. Both Ströbel et al. (2018) and Kulkarni et al. (2018) used machine learning to categorise text. While Ströbel et al. (2018) categorised by genre, Kulkarni et al. (2018) analysed the contents of books to predict publication dates.

3.0 Time period categorisation

"Time" by Barbara Adam (2006) is an interdisciplinary book that delves into the complex concept of time, bringing together insights from various fields such as philosophy, sociology, and anthropology. Time is experienced, understood and valued differently across cultures, and it has a multifaceted nature that has been conceptualised and measured throughout history (Adam 2006). Time period is a concept for conceptualising and measuring time. There are well-known historical periods that refer to wars, revolutions and inventions, such as the Medieval period, the Viking Age and World War I. However, Adam (2006) points out that naming a period often occurs after the event has taken place, and there are no strict guidelines governing what qualifies as a time period.

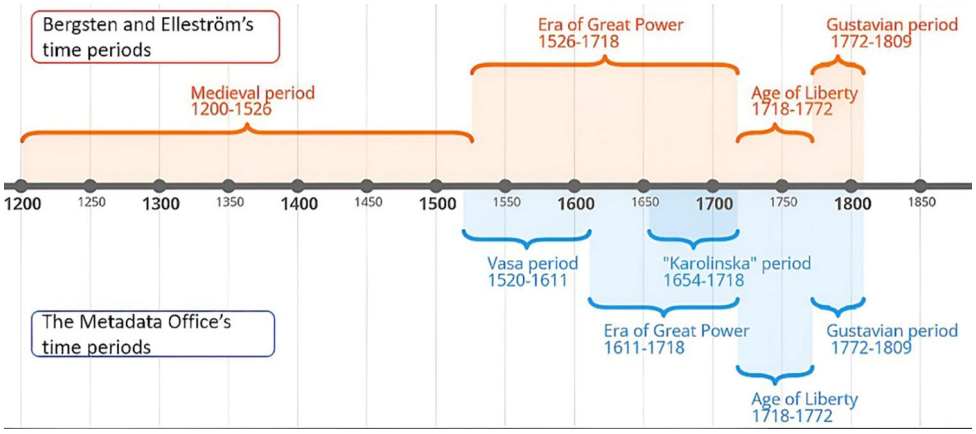

Figure 1. Time periods from the Metadata Office and Bergsten and Elleström (2004).

There are many different types of categories that can be used to categorise time. The most basic type of categorisation, according to Frommeyer (2013), is by identifying periods and events in the order of occurrence in a particular year or decade. Other types of categorisations are technological categorisation, for example, in relation to the Industrial Revolution or the Information Age (Britannica 2022), and economic and political categorisation, for example, in relation to the Great Depression or the rise of capitalism and the type of government in power (Bergsten and Elleström 2004). Bergsten and Elleström (2004) argue that because time can be categorised in numerous ways, it can be challenging and contentious for scientists to agree on conceptualising and determining the start and end points of time periods. Also, Shaw (2010) points out that time periods are concepts of human thought, and like any concept, they can change with the occurrence of new events or interpretations. Since this study is mainly preoccupied with Swedish language literature, the focus will be on Nordic time periods. This study utilises the division of time periods proposed by Bergsten and Elleström (2004), which is primarily aligned with the period divisions established by the Metadata Office (2021a). The Metadata Office, part of the National Library of Sweden, creates metadata standards and cataloguing instructions that libraries widely use to generate time period metadata during the cataloguing process. These time periods are mostly political periods, except for the Medieval period, and they undergo revisions and refinements over time. Figure 1 shows Bergsten and Elleström's time periods and the time periods provided by the Metadata Office. Upon examination, it becomes clear that there are some differences between the two. The most significant difference is that the Medieval Period, which is a long time period, is not included in the Metadata Office's list. Another difference is that Bergsten and Elleström's periods span longer periods, while the Metadata Office's time periods are shorter and more specific. A detailed description of how they are joined and used in this study is provided in the Data section.

4.0 Method

Quasi-experiments were used in this study to categorise Swedish historical fiction texts using three machine learning techniques. The efficiency of these techniques is measured using the F1-score.

According to Cook (2015, 1-2), a quasi-experiment "aims to establish a cause-and-effect relationship between an independent and dependent variable". In contrast to so-called true experiments, Cook (2015, 1-2) further notes that it "does not rely on random assignment. Instead, subjects are assigned to groups based on non-random criteria". In this study, each novel is assigned a time period, which are non-random groups. The research employs machine learning techniques to create features and categorise texts using supervised and unsupervised learning methods. In supervised learning, the algorithm is provided with inputs and known outputs to learn how to categorise texts, while unsupervised learning identifies patterns without known outputs (Burkov 2019). Both techniques are utilised in this research since there are known and unknown variables. Unsupervised learning is used to create text features without human supervision, while supervised learning verifies whether the predicted time period categories are correct or wrong.

4.1 Data preparation

This research utilised historical fiction novels from the Swedish Literature Bank, a nonprofit initiative to offer free access to digital versions of Swedish literature. Initially, 48 novels were searched and retrieved in full text by submitting the phrase "historical novels" to the search interface. Historical fiction novels were chosen due to their basis on historical

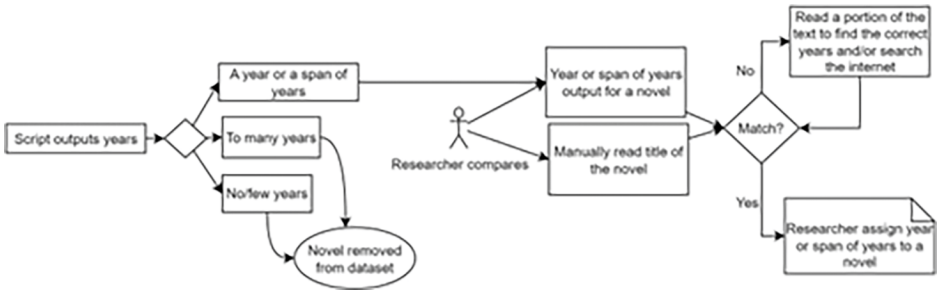

Figure 2. Manual historical fiction novel categorisation by time periods.

events, making them easier to categorise into a time period than other novels. Since the novels did not have time period data, they had to be manually categorised, resulting in a dataset of 35 novels, which is considered a small dataset for ma chine learning. The novels were also lengthy, ranging from 38,000 to 186,000 words. To address these issues, the novels were sliced into smaller chunks, resulting in 1,055 individual texts with around 3,500 words each. Slicing the novels into smaller pieces helped increase the dataset size and made the texts more manageable for machine learning. When I mention 'novels', I refer to complete, unaltered literary works. On the other hand, when I use the term 'text', it denotes the novels that have been divided into smaller segments or sections. The distinction is important since feature engineering and categorisation are done on smaller text segments, not the whole novel.

The Metadata Office (2021) and Bergsten and Elleström (2004) offer suitable time period categories for literature categorisation. However, time periods by Bergsten and Elleström and the Metadata Office were slightly revised due to the data used in this study. There was a considerable number of novels with Medieval time periods, therefore this period was added as a supplementary category, as suggested by Bergsten and Elleström. The Vasa period was excluded from the list of time period categories since there was only one novel set at that time. The Karolinska period was merged with the Era of Great Power because subcategories were not allowed. Additionally, different periods ended and started in the same year, creating ambiguity. To avoid this, one period had to end a year before the next period started. After the revisions, the time period categories are as follows: 1200-1520, Medieval period; 1611-1717, Era of Great Power; 1718-1772, Age of Liberty; and 1773-1809, Gustavian Period. A script was created to scan each novel for any numerical years mentioned. These years were important in determining when in time the story takes place. The years extracted from each novel were individually examined to select a time period. If a novel had no years mentioned, it was excluded from the dataset. However, if a novel had too many different years mentioned (30 or more), and none stood out, then they were also excluded to avoid guesswork. Historical fiction novels usually have titles that mention important details like the names of kings, queens, wars, or years, and as an additional step, the novel's titles were compared to the extracted years to increase the accuracy of the time period categorisation. After the manual categorisation, each time period category had 100 texts each. The steps of manual time period categorisation are shown in Figure 2.

4.2 Text pre-processing

In order to make sure that the texts were consistent for analysis, several steps were taken during pre-processing. First, all the texts were converted to lowercase to make sure that words that meant the same thing, like "King" and "king", were treated as identical. The texts were also divided into smaller parts, or tokens, by splitting them into individual words. Some words and punctuation marks that did not add to the interpretation of the text, such as "I," "and," "she," ".", ":" and "?" were removed. The texts for LDA and TF-IDF were cleaned, but for WE with SBERT, the texts remained as they were. This is because SBERT works on the sentence level, and splitting the words or removing punctuation makes it challenging to distinguish sentences apart.

4.3 Feature engineering

After the pre-processing stage, the texts were transformed from whole texts into a sequence of chosen words. However, machine learning algorithms cannot directly use words in this format; thus, the words need to be converted into numerical values. To achieve this, each text is represented as a list of numbers, also known as feature vectors. This study used three different techniques to create these feature vectors: LDA, WE, and TF-IDF. LDA is a type of generative probabilistic model in the topic modeling family that aims to identify the set of topics present in a document. LDA assumes that a document contains words corresponding to various topics and assigns a set of topic probabilities to each document. These probabilities range from 0 to 1, with 1 indicating a strong probability that the topic exists in the document and 0 indicating a low probability (Blei, Ng and Jor-

dan, 2003). Various topic numbers were tested to find the optimal number of topics. Using fewer than five topics gave poor results, which caused the algorithm to incorrectly categorise a text 50 percent of the time, which was as good as a random guess. However, as the number of topics increased, there was a clear improvement in the results. For example, using ten topics was better than using 5, and using 15 topics was better than using 10. This pattern continued up to 20 topics, after which there were minimal improvements. Beyond 20 topics, the results worsened, dropping from 79 percent to 78 percent in correctly predicting a text's time period. This suggests that meaningful topics could be formed with 20 topics, therefore 20 topics are used in this study.

TF-IDF is a statistical model that measures a word's significance in a document or a collection of documents. It does this by calculating the frequency of the word in the document and weighting it according to how common or uncommon the word is across all documents (Shalev-Shwartz and Ben-David 2014). The model uses two measures: term frequency (TF) and inverse document frequency (IDF). TF counts the number of times a word appears in a specific document, while IDF measures the rarity of the word across all documents. A high document frequency means a lower IDF score, while a low document frequency means a higher IDF score. Words that frequently appear in 90 percent of the documents were removed because they carry little meaning and do not provide information to differentiate one document from another. Similarly, words that are too rare (appear in less than 10 percent of the documents) or unique were also removed because they are unlikely to be useful in distinguishing between documents. Misspelled words or proper names that appear only in a single document were also unlikely to help identify relevant documents.

In natural language processing, word embedding represents words as numerical vectors in a high-dimensional space. SBERT is a type of pre-trained model that is available for generating sentence embeddings (Reimers and Gurevych 2019). The training aims to maximise the similarity between the sentence embeddings of semantically similar sentence pairs and minimise the similarity of semantically different sentence pairs. When given a sentence as input, SBERT generates a fixed-length vector representation of the sentence called sentence embedding, which captures the semantic meaning of the sentence input. This is achieved by encoding the input sentence into a sequence of token embeddings using the pre-trained BERT model. Unlike LDA and TF-IDF, the data were not pre-processed for SBERT because punctuation is needed to distinguish between sentences.

In this feature engineering step, the texts were transformed into feature vectors, which are numerical representations of the text. Each algorithm has produced its representations by analysing the text from different perspectives.

4.4 Model training

Logistic Regression is a model used for categorical dependent variables, specifically for binary categorisation problems. In this study, there are four categories of time periods, which creates a multi-category categorisation problem. The approach to solving this problem is called "one vs all", where one category is compared to the remaining combined categories. The logistic regression model uses feature vectors to predict whether a text belongs to category A by outputting a 1 or 0. Since Logistic Regression is a supervised algorithm, pre-labelled data is needed to train it. In this case, each text was manually labelled with its correct time period before the model training. K-fold cross-validation with ten folds was used to train and test the LDA, TF-IDF, and WE with SBERT models. The output from this step is a categorisation model that predicts the probability of a text belonging to a certain category.

4.5 Model evaluation

The three models, LDA, TF-IDF, and WE with SBERT, were evaluated separately and compared using the F1-score, a common accuracy measure of categorisation models. The F1-score is the harmonic mean of precision and recall, with a maximum value of 1.0 indicating perfect precision and recall. The minimum value of 0 indicates that the categorisation models did not identify any true positives correctly. When dealing with multiple categories, an overall F1 score is not computed. Instead, a one-vs-all scoring method determines the F1 score for each category. This method assesses the performance of each category individually, using precision and recall. In categorisation tasks, precision and recall are two metrics commonly used to evaluate the performance of a model. The F1 score is a metric that combines precision and recall into a single score that reflects the model's overall performance. Precision is a metric that measures the proportion of true positives out of all positive predictions the model makes, see Equation 1. Precision measures how many instances the model predicted as positive are positive. A high precision score means the model makes very few false positive predictions. Recall is a metric that measures the proportion of true positives from all actual positive instances in the dataset, see Equation 2. Recall measures how many positive instances in the dataset are correctly identified by the model. A high recall score means that the model correctly identifies many positive instances. The F1 score is a weighted average of precision and recall, with equal weight given to both metrics, see Equation 3. A high F1 score means the model has high precision and recall, which is desirable in many categorisation tasks.

$$\text{Precision} = \frac{T_p}{T_p + F_p} \quad (1)$$

$$\text{Recall} = \frac{T_p}{T_p + T_n} \quad (2)$$

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

5.0 Results

The results of each technique are presented to show how they performed through precision, recall, and F1-score for each time period category. Lastly, in 5.1, a comparison between techniques is presented.

Figure 3 shows how the technique LDA performed across the time period categories. Out of the four time period categories, Age of Liberty had the highest precision, recall and F1 scores, followed by the Medieval period. However, both the Era of Great Power and the Gustavian period had less accurate predictions of 0,74, which equals correct predictions 74 percent of the time.

Figure 4 shows that TF-IDF has high scores on precision, recall and F1-score across all time period categories. However, the Medieval period and Age of Liberty scored highest, followed by the Era of Great Power and the Gustavian period.

In contrast to TF-IDF, Figure 5 shows that WE SBERT has a low score on all metrics. The medieval period and the

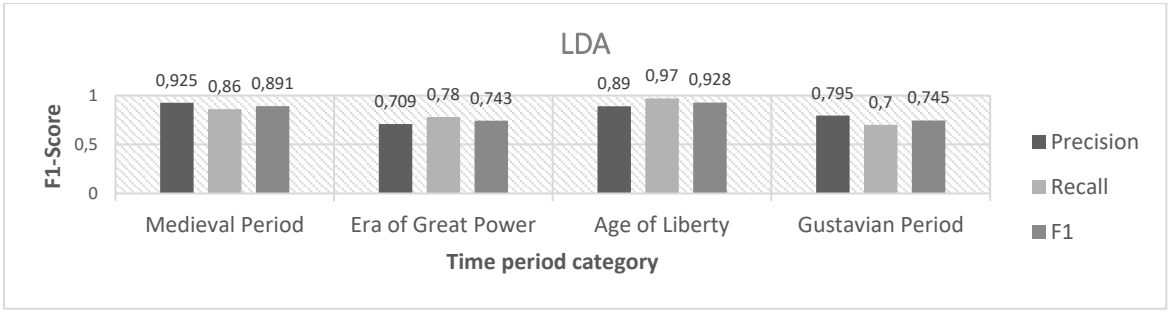

Figure 3. LDA F1-score for each time period.

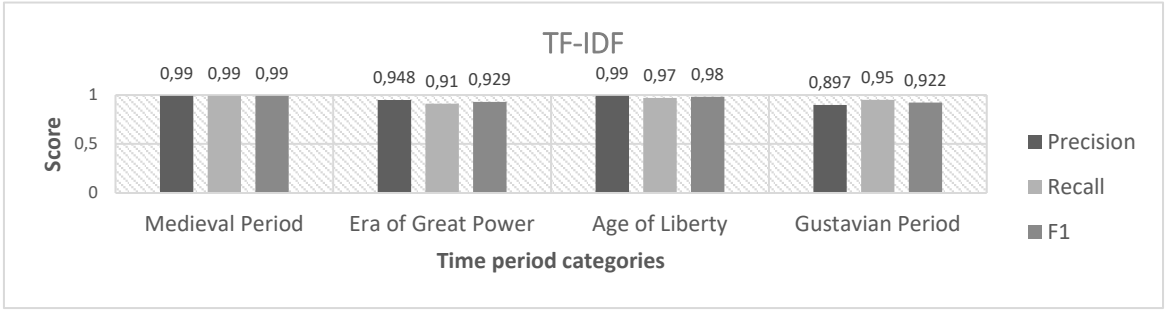

Figure 4. TF-IDF F1-score for each time period.

Figure 5. WE SBERT F1-score for each time period.

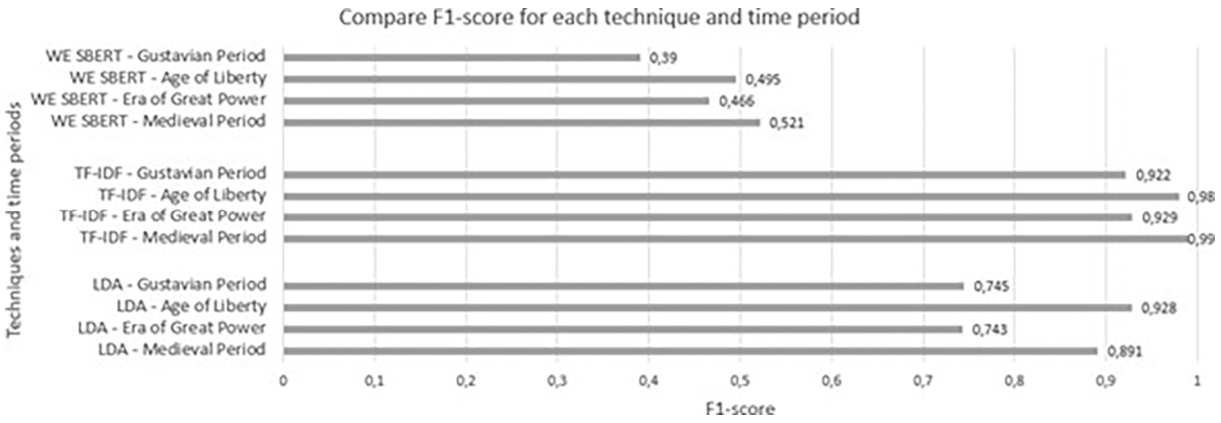

Figure 6. F1-score comparison between techniques for each time period category.

Age of Liberty have scores close to 0.5, meaning that only 50 percent of the texts were categorised correctly. The era of Great Power and the Gustavian period had even lower F1 scores, meaning that more text was categorised wrong than correct.

5.1 Comparison between techniques with F1-score

Figure 6 shows the F1-score comparison between the three techniques. The data shows that TF-IDF performed well across all four time periods, with scores between 0.92 and 0.99. LDA also performed well, with scores ranging between0,74 and 0.92. WE SBERT had the lowest scores overall, with scores ranging from 0.39 to 0.521. Regarding the specific time periods, the Age of Liberty has the highest F1 scores across all three techniques, while the Gustavian Period has the lowest scores.

6.0 Discussion

The results improved as the quasi-experiments proceeded. The study results suggest that TF-IDF and LDA are promising feature engineering techniques for analysing text data across different time periods. At the same time, WE SBERT may be less effective in this context. LDA and TF-IDF performed consistently well across all four time periods, with scores ranging from 0.74 to 0.99, while WE SBERT had lower scores, about and under 0.5, across all four periods. Several factors may contribute to the differences in performance between these techniques. For example, LDA and TF-IDF are well-established techniques widely used in natural language processing and may be better suited to analysing texts from a range of time periods. Additionally, LDA and TF-IDF rely on statistical methods to identify patterns and trends in text data, which may be particularly useful for identifying similarities and differences across periods. On the other hand, WE SBERT is a newer technique that uses neural network models to generate vector representations of sentences. However, using too many sentences may result in lower-quality embeddings, which affect the categorisation.

As Saarti (2019) discussed, fiction content analysis is challenging due to its multimodality and interpretational nature. However, as this study's results suggest, machine learning algorithms can be used effectively to find patterns and relationships in large amounts of data and categorise features of historical fiction texts.

Fiction categorisation has traditionally relied on formal and external aspects such as genre, literary form, author, place, and language. However, as pointed out by Almeida and Gnoli (2021) and as corroborated by the results of the present study, a more effective approach to categorising fiction is to consider the actual content of the texts. Doing so makes it possible to capture the thematic essence of fiction texts and provide a more accurate categorisation.

Additionally, as machine learning techniques become more available, areas of fiction categorisation that were previously explored manually can now be explored with ML, such as identifying if a text is fiction or nonfiction, as done by Manger (2018) or, more commonly, categorising based on genre or publication dates (Ströbel et al.2018 and Kulkarni et al. 2018). To improve the categorisation of fiction, it is necessary not only to reassess traditional categories like genre and publication date with machine learning but also to consider less traditional approaches for categorising fiction, such as time period categorisation. Time is not a new phenomenon when organising literature in libraries, but only a small fraction of literature has been categorised in this way.

The findings in this study have several implications for future research and practical applications. First, researchers and practitioners interested in analysing text data across different time periods may benefit from using LDA or TF-IDF as feature engineering techniques, as these techniques have consistently been performed across all time periods. Second,

more experiments should be done to optimise WE SBERT for time period categorisation, i.e., to vary the number of input sentences to get higher performance. It is important to note that this study had some limitations, including the small sample size and insufficient data to cover the Vasa period. Future research could explore the performance of these techniques on more extensive or diverse datasets with more time periods, and investigate different pre-processing and categorisation algorithms. Moreover, this study explored one-to-one categorisation, meaning that one text could only belong to one time period, whereas future research could explore multiple possible categories for each given text.

7.0 Conclusion

The aim of this study was to compare different feature engineering techniques alongside preprocessing and classification. Furthermore, the aim was to evaluate the models with F1 score, precision, and recall, determining the most accurate approach for predicting time periods in Swedish historical fiction literature sourced from the Swedish Literature Bank. The evaluation utilised the metrics F1 score, precision, and recall. The categories for prediction included the Medieval Period, Era of Great Power, Age of Liberty, and Gustavian Period. During the preprocessing stage, it was observed that Swedish text presented challenges compared to English, especially with stemming. Initially, all texts underwent preprocessing, but it was noted that SBERT yielded unsatisfactory results with preprocessing. As a result, only LDA and TF-IDF were pre-processed. Preprocessing also presented challenges due to the use of the old Swedish language, making it difficult to remove most of the stop words. In terms of feature engineering, quasi-experiments revealed varying performances among the techniques. Overall, the results suggest that TF-IDF and LDA are promising techniques for categorising text data across different time periods. At the same time, WE with SBERT produced poor results for all three time periods. TF-IDF and LDA exhibited better performance, with F1 scores ranging between 0.74 and 0.99 across the four time periods, showcasing their effectiveness in capturing temporal patterns within the text data. This indicates that TF-IDF and LDA accurately classified texts between 74 and 99 percent of the time.

Conversely, WE SBERT yielded lower scores, generally falling below the threshold of 0.5 for all time periods, indicating its limited suitability for time period categorisation in this context. In this study, only one classification algorithm, namely logistic regression, was utilised. Future research could explore the effectiveness of other classification algorithms and assess whether they yield improved results, particularly concerning SBERT.

References

Adam, B. (2004). *Time*. Cambridge and Malden, MA: Polity Press

Almeida, Patricia de, and Claudio Gnoli. 2021. "Fiction in a Phenomenon-Based Classification". *Cataloging & Classification Quarterly* 59 no.5: 477-91. https://doi.org/10.1080/01639374.2021.1946232

Bates, Marcia J., Deborah. N. Wilde, and Susan. Siegfried. 1993. "An Analysis of Search Terminology Used by Humanities Scholars: The Getty Online Searching Project Report Number 1". *The Library Quarterly* 63 no.1: 1-39.

Bergsten, Stafan and Lars Elleström. 2004. *Litteraturhistoriens Grundbegrepp*. 2nd. ed. Studentlitteratur *AB* Lund, Sweden.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3 no.1: 993-1022.

Bogaard, Tessel, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. 2019. "Metadata Categorization for Identifying Search Patterns in a Digital Library". *Journal of Documentation* 75 no.2: 270-286.

Britannica, History of Technology, 25 Aug. 2022. https://www.britannica.com/technology/history-of-technology/Military-technology.

Burkov, Andriy. 2019. *The Hundred-page Machine Learning Book.* Vol.1. Quebec City, Canada.

Cook, Thomas D, ed. 2015. "Quasi-experimental Design". In *Wiley Encyclopedia of Management* 1-2.

Dalli, Angelo. 2006. "Temporal Classification of Text And Automatic Document Dating". In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, Association for Computational Linguistics*, New York, USA, 29-33. doi.10.3115/1629235.1629238

Frommeyer, Jutta. 2013. "Chronological Terms and Period Subdivisions in LCSH, RAMEAU, and RSWK". *Library Resources & Technical Services* 48 no.3:199-212.

Guerrini, Mauro.2023. *From Cataloguing to Metadata Creation: a Cultural and Methodological Introduction*. Facet Publishing.

Gartner, R., L'Hours, H. and Young, G., 2008. *Metadata for Digital Libraries: State of the Art and Future Directions*. JISC.

Håkansson, Anna, and Ronald Lee Hartung. 2020. *Artificial Intelligence: Concepts, Areas, Techniques And Applications*.1st. ed. Studentlitteratur AB, Lund, Sweden.

Kulkarni, Vivek and Yingtao Tian, Parth Dandiwala, and Steve Skiena. 2018. "Simple Neologism Based Domain Independent Models to Predict Year of Authorship". *In Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics*, Santa Fe, New Mexico, USA, 202-212.

Maker, Richard. 2008. "Finding What You're Looking For: A Reader-Centred Approach To The Classification Of Adult Fiction In Public Libraries." *The Australian Library Journal* 57 no.2: 169-77.

Manger, Clement. 2018. "That Seems Made Up: Deep Learning Classifiers for Fiction & Non-Fiction Book Reviews". Masters' diss., Technological University Dublin, Republic of Ireland.

Metadatabyrån (Metadata Office) 2021a. "Lista kronologiska ämnesord särskilda länder". 24 Mar. 2021, URL: https://metadatabyran.kb.se/amnesord-och-genre-form/svenska-amnesord/typer-av-amnesord/kronologiska-amnesord/lista-allmanna-kronologiska-amnesord.

Metadatabyrån (Metadata Office) 2021b. "Principer för ämnesordsindexering". 24 Mar. 2021, URL: metadatabyran.kb.se/amnesord-och-genre-form/svenska-amnesord/typer-av-mnesord/kronologiska-amnesord/lista-kronologiska-amnesord-sarskilda-lander.

Petras, V., Larson, R. R., & Buckland, M. 2006. "Time period directories: a metadata infrastructure for placing events in temporal and geographic context", In Proceedings of the 6th acm/ieee-cs joint conference on digital libraries, 151–160.

Rafferty, Pauline. 2013. "Epistemology, Literary Genre and Knowledge Organization Systems". In *Actas del X Congreso de ISKO-España. Ferrol 20 de junio-1 de julio de 2011,* Ferrol, Spain, edited by Mª Carmen Pérez Pais and María G. Bonome. Ferrol : Universidade da Coruña, Servizo de Publicacións, 553-565.

Reimers, Niels and Gurevych, Irina. 2019. "Sentence-bert: Sentence Embeddings Using Siamese Bert-Networks". Preprint. submitted August 27, 2019. *arXiv:1908.10084*. doi.org/10.48550/arXiv.1908.10084.

Saarti, Jarmo. 2019. "Fictional Literature, Classification and Indexing". *Knowledge Organization* 46 no.4: 320-332. doi.org/10.5771/0943-7444-2019-4-320

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory To Algorithms*.1st. ed. Cambridge University Press.

Shaw, Ryan. 2010. *Events and Periods as Concepts for Organizing Historical Knowledge*. 1st ed. University of California, Berkeley.

Shenton, Andrew K. 2006. "The Role of 'Reactive Classification' in Relation to Fiction Collections in School Libraries". *New Review of Children's Literature and Librarianship* 12 no.2: 127-46.

Ströbel, Marcus, Elma Kerz, Daniel Wiechmann and Yu Qiao. 2018. "Text Genre Classification Based on Linguistic Complexity Contours Using a Recurrent Neural Network". In *Proceedings of the Tenth International Workshop Modelling and Reasoning in Context* Stockholm, Sweden, 56-63.