Comparing Feature Engineering Techniques for the Time Period Categorization of Novels

Fereshta Westin[1]

[1] *University of Borås, Allégatan 1, Borås, Sweden*

Abstract

Categorizing literary works based on the time period in which the story is set can enhance the accessibility of these works to library patrons through online catalogues. However, this categorization is not frequently implemented, which makes it difficult for patrons to find works based on their historical context. The reason is that assigning time period categories is a tedious and time-consuming task that requires extensive analysis by librarians and cataloguers. To address this issue, this paper suggests using machine learning techniques, Latent Dirichlet Allocation (LDA), Term Frequency-Inverse Document Frequency (TF-IDF), and Word Embedding (WE) with Sentence-BERT (SBERT) to categorize literary works by time periods automatically. LDA identifies underlying topics within the text, TF-IDF measures word importance, and SBERT measures the semantic similarity between sentences, enhancing the precision of categorization. The study evaluates the accuracy of these different techniques using a dataset of Swedish historical fiction pieces from the Swedish Literature Bank while performing quasi-experiments. Based on the F1-score, the study found that TF-IDF and LDA are effective techniques for categorizing text data by time period. On the other hand, the results indicate that WE with SBERT did not perform well for any of the three time periods analyzed.

Keywords

Literary categorization, machine learning techniques, time period categorization

1. Introduction

One way to organize literature is by including metadata about specific time periods. Frommeyer [1] argues that time is an important component of subject cataloging, and a study by Bates et al. [2] found that 16% of humanities scholars' searches on the DIALOG database were related to time. In this study, "time" referred to the time period in which the story takes place (e.g. the Medieval period, the Viking age, or World War I), not the publication date of the literary work. Metadata that represents time periods can be added to search engines in online collections, such as library systems, to increase the chances that users can find the literature they are looking for [3].

The National Library of Sweden and its Metadata Office are responsible for maintaining the guidelines for categorizing literary works by time periods; educated cataloguers at different libraries are responsible for this categorization. In Sweden, the joint library catalogue (Libris) which is used by cataloguers has around nine million printed works of fiction and two million printed works of non-fiction. A search in Libris database shows that out of the 12 million works, 108,816 works had a dedicated time period specified. These numbers support the claim made in Dalli's [4] study that it is rare to find literary works with time period metadata, making time period searches difficult.

Although it is possible to categorize literary works by time periods, it is often not prioritized, 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 on what metadata to specify in each case. Cataloguers categorizing in Libris can obtain time period information from the author, publishers, other cataloguers, or organizations. In the absence of time period information, cataloguers must read the blurb or a portion

ISKO UK conference, July 24–25, **20**23, Glasgow, Scotland
EMAIL: email1@mail.com (A. 1); email2@mail.com (A. 2); email3@mail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)

of the text, search the internet or ask colleagues. Understanding 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-categorization. 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% of its content dedicated to describing one or more time periods [5]. In the process of cataloguing, the work of a cataloguer involves understanding a work's characteristics and categorizing it accurately, which is similar to feature engineering tasks in Machine Learning (ML).

Understanding text characteristics is known as feature engineering, which is a common task in ML used to identify patterns in data and make predictions [6]. As with a cataloguer, the effectiveness of a categorization is heavily influenced by the quality of the text analysis.

In this study, the aim is to evaluate and compare different feature engineering techniques to determine which one analyses the data most accurately leading to accurate predictions of time periods for Swedish historical fiction literature from the Swedish Literature Bank. By using ML techniques to aid cataloguers in their decision-making process, this study could improve the accuracy and efficiency of time period categorization. The 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), and they are evaluated by F1 score. More information is provided in the Methods section.

2. Fiction categorization

The topic of fiction content analysis and retrieval has become increasingly significant in the context of knowledge management and literature organization, as pointed out by Saarti [7]. However, categorizing fiction has proven challenging due to its multifaceted and interpretive nature [7]. Rafferty [8] observes that genre has traditionally been employed to categorize fiction, owing to its usage in advertising and targeted marketing [7, 9]. Shenton [10] 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 [11] claim that conventional categorization systems which index fictional works based only on their form, genre, and language may not be the most effective approach, since they need to consider the actual content of the story. Therefore, it is essential to develop new methods to better analyse and categorize fiction content, as traditional categorization systems have been found to be inadequate in this regard.

There are numerous ways of categorizing fiction, as evidenced by various attempts. However, one prominent approach involves analysing the content. ML can simplify this task by enhancing the ease of conducting content analysis. Manger [12] explored the possibilities of using ML to categorize a text as fiction or nonfiction by analysing reviews. Both Ströbel et al. [13] and Kulkarni et al. [14] used ML to categorize text. While Ströbel et al. [13] categorized on genre, Kulkarni et al. [14] analysed the contents of books to predict publication dates.

3. Time periods categorisation

"Time" by Barbara Adam [15] 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 conceptualized and measured throughout history [15]. Time period is a concept for conceptualizing 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, naming a period often occurs after the event has taken place, and there are no strict guidelines governing what qualifies as a time period [15].

There are many different types of categories that can be used to categorize time. The most basic type of categorization is by identifying periods and events in the order of occurrence in a particular year or decade [1]. Other types of categorizations are technological categorization for example in relation to the Industrial Revolution or the Information Age [16], and economic and political categorization for example in relation to the Great Depression or the rise of capitalism and the type of government in power [17].

Because time can be categorized in numerous ways, it can be challenging and contentious for scientists to agree on how to conceptualize and determine the start and end points of time periods [17]. Also, time periods are concepts of human thought, and like any concept, they can change with the occurrence of new events or interpretations [18]. Since this study mainly is preoccupied with Swedish language literature, the focus will be on Nordic time periods. This study utilizes the division of time periods proposed by Staffan Bergsten and Lars Elleström [17], which is largely aligned with the period divisions established by the Metadata Office [20]. 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 Medieval period, and they\ undergo revisions and refinements over time. Figure 1 shows Bergsten and Elleström's time periods and those 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.

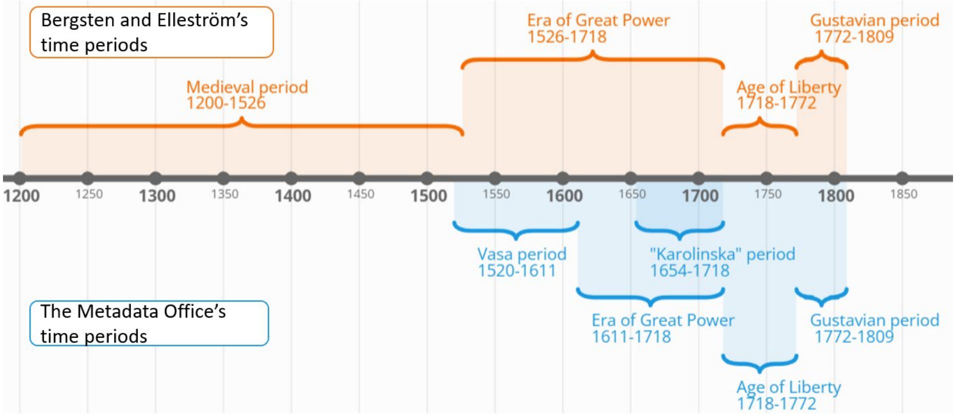
Bergsten and Elleström's time periods
Era of Great Power 1526-1718
Gustavian period 1772-1809
Medieval period 1200-1526
Age of Liberty 1718-1772
1200
1250
1300
1350
1400
1450
1500
1550
1600
1650
1700
1750
1800
1850
Vasa period 1520-1611
"Karolinska" period 1654-1718
The Metadata Office's time periods
Era of Great Power 1611-1718
Gustavian period 1772-1809
Age of Liberty 1718-1772

Figure 1: Time periods from the Metadata Office and, Bergsten and Elleström

4. Method

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

According to Cook [19], 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 [19] further notes, 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 ML techniques to create features and categorize texts using supervised and unsupervised learning methods. In supervised learning, the algorithm is provided with inputs and known outputs to learn how to categorize texts, while unsupervised learning identifies patterns without known outputs [21]. Both techniques are utilized in the research since there are known and unknown variables. Unsupervised learning is used to create text features without human supervision, while supervised learning is used to verify if the predicted time period categories are correct or wrong.

4.1 Data preparation

This research utilized 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 events, making them easier to categorize into a time period than other novels. Since the novels did not have time period data, they had to be manually categorized, resulting in a dataset of 35 novels, which is considered a small dataset for ML. 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 ML. When I mention 'novels', I am referring 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 categorization is done on the smaller segments of text and not the whole novel.

The Metadata Office and Bergsten and Elleström offer suitable time period categories for literature categorization. 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 in 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 name of kings, queens, wars, or years, and as an additional step, the novels titles were compared to the extracted years to increase the accuracy of the time period categorization. After the manual categorization, each time period category had 100 texts each. The steps of manual time period categorization are shown in Figure 2.

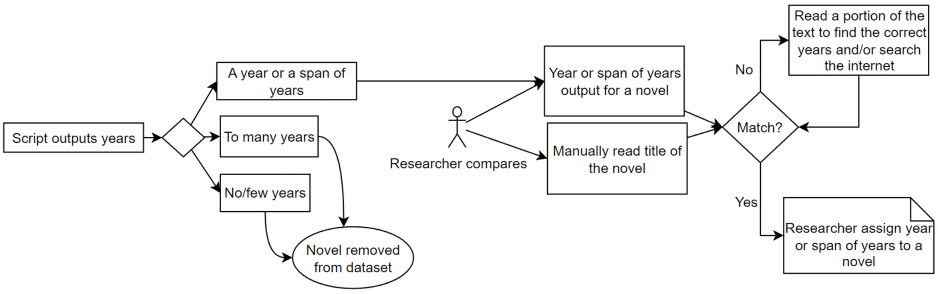
Script outputs years
A year or a span of years
To many years
No/few years
Novel removed from dataset
Researcher compares
Year or span of years output for a novel
Manually read title of the novel
Match?
No
Read a portion of the text to find the correct years and/or search the internet
Yes
Researcher assign year or span of years to a novel

Figure 2: Manual historical fiction novel categorization by time periods

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. Additionally, the texts were broken up into smaller parts, or tokens, by splitting it into individual words. Some words and punctuation marks that didn't 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, ML 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 modelling family that aims to identify the set of topics present in a document. LDA assumes that a document contains words that correspond 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 [22]. To find the optimal number of topics, various topic numbers were tested. Using fewer than five topics gave poor results, which caused the algorithm to incorrectly categorize a text in 50% 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% to 78% in correctly predicting a text's time period. This suggests that meaningful topics could be formed with 20 topics, which 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 [24]. 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 appear frequently in 90% 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% of the documents) or unique were also removed because they are unlikely to be useful in distinguishing between documents. Misspelt 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 available for generating sentence embeddings [23]. The training aims to maximize the similarity between the sentence embeddings of semantically similar sentence pairs, and minimize 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 preprocessed 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 analyzing the text from different perspectives.

4.4 Model training

Logistic Regression is a model used for categorical dependent variables, specifically for binary categorization problems. In this study, there are four categories of time periods, which creates a multi-category categorization 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 first manually labelled with its correct time period before the model training. K-fold cross-validation with 10 folds was used to train and test the LDA, TF-IDF, and WE with SBERT models. The output from this step is a categorization 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 categorization 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 categorization 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 categorization 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 categorization tasks.

$$Precision = \frac{T_p}{T_p + F_p} \tag{1}$$

$$Recall = \frac{T_p}{T_p + T_n} \tag{2}$$

$$F1 = \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

5. Results

The results are presented for each technique 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 Medieval period. However, both Era of Great power and Gustavian period had less accurate predictions of 0,74, which equals to correct predictions in 74% of the time.

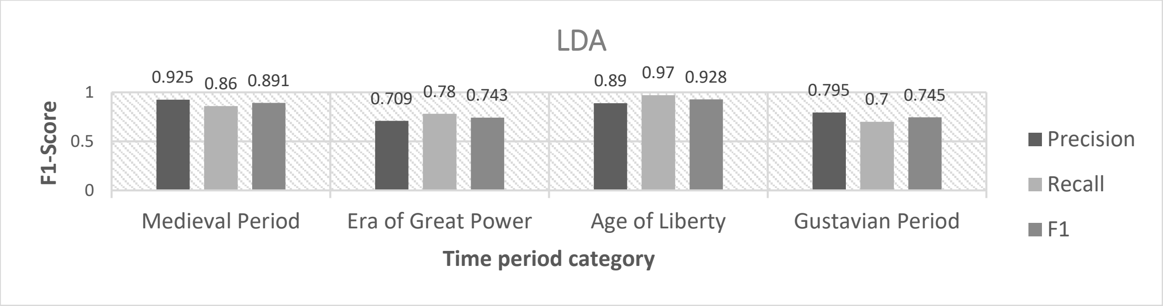

Figure 3: LDA F1-score for each time period

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

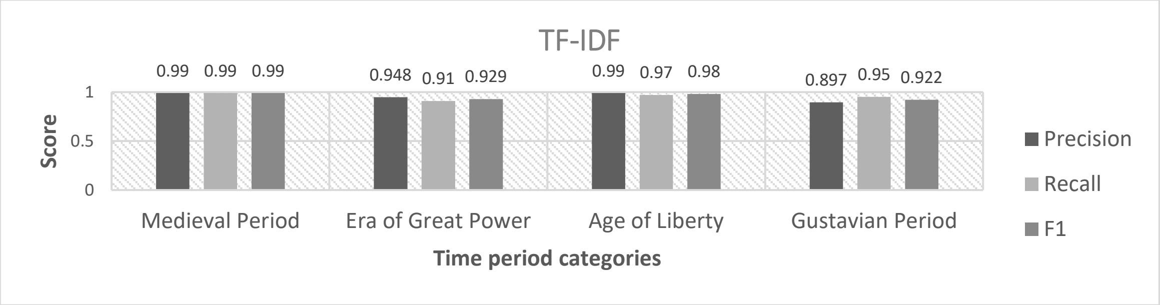

Figure 4: TF-IDF F1-score for each time period

As a contrast to TF-IDF, Figure 5 shows that WE SBERT has low score on all metrices. Medieval period and Age of Liberty have scores close to 0,5 which means that only 50% of the texts were categorized correctly. Era of Great power and Gustavian period have even lower F1-score and it means that more text were categorized wrong than correct.

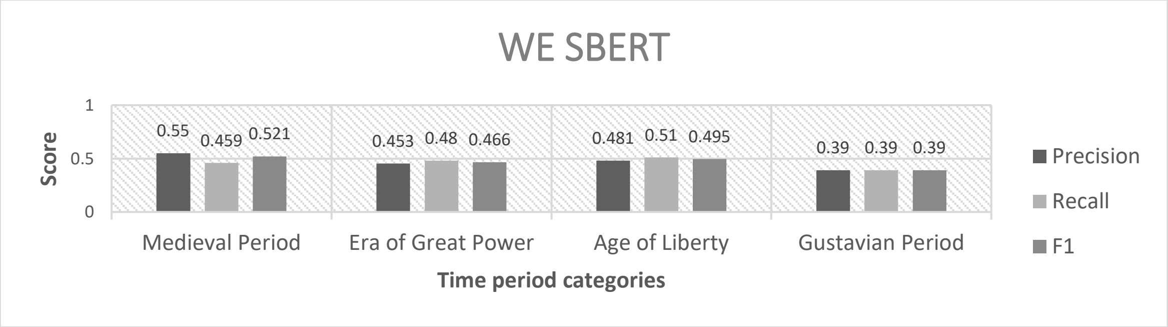

Figure 5: WE SBERT F1-score for each time period

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 ranging 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. In terms of 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.

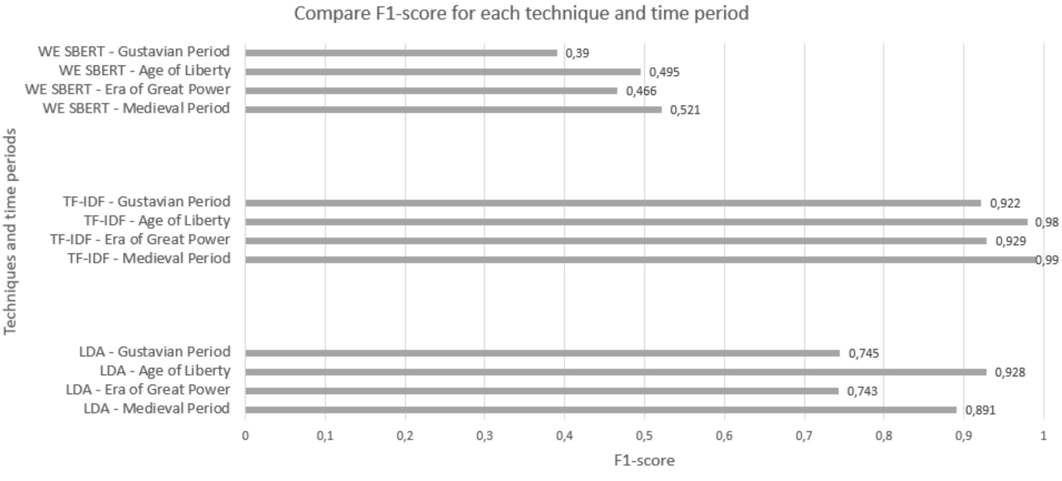

Figure 6: F1-score comparison between techniques for each time period category

6. Discussion

The results improved as the quasi-experiments proceeded. The study results suggest that TF-IDF and LDA are promising feature engineering techniques for analyzing text data across different time periods, while WE SBERT may be less effective in this context. Both 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 across all four periods, about and under 0.5. 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 analyzing texts from a range of time periods. Additionally, LDA and TF-IDF rely on statistical methods that can 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 categorization.

As Saarti [7] discussed, fiction content analysis is challenging due to its multimodality and interpretational nature. However, as the results of this study suggest, ML algorithms can be used effectively for analysis to find patterns and relationships in large amounts of data and categorize features of historical fiction texts.

The categorization of fiction 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 [11], and as corroborated by the results of my study, a more effective approach to categorizing fiction is to consider the actual content of the texts. By doing so, it becomes possible to capture the thematic essence of fiction texts and provide a more accurate categorization.

Additionally, as ML techniques getting more available, areas of fiction categorization 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 [12], or, more commonly, categorizing based on genre or publication dates [13,14]. To improve the categorization of fiction, it is necessary not only to reassess traditional categories like genre and publication date with ML but also to consider less traditional approaches for categorizing fiction, such as time period categorization. Time is not a new phenomenon when organizing literature in libraries, but only a small fraction of literature has been categorized in this way.

The findings in this study have several implications for future research and practical applications. First, researchers and practitioners interested in analyzing text data across different time periods may benefit from using LDA or TF-IDF as feature engineering techniques, as these techniques have shown consistent performance across all time periods. Second, more experiments should be done to optimize WE SBERT for time period categorization, 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 not enough data to cover the Vasa period. Future research could explore the performance of these techniques on larger or more diverse datasets with more time periods, and investigate different pre-processing and categorization algorithms. Moreover, this study explored one-to-one categorization, meaning that one text could only belong to one time period, whereas future research could explore multiple possible categories for each given text.

7. Conclusion

This study aimed to evaluate and compare three feature engineering techniques to discover which technique excels in engineering features, with the aim of categorizing historical fiction novels by the correct time period: Medieval Period, Era of Great Power, Age of Liberty and Gustavian Period. This was done by experimenting with the following ML techniques: LDA, TF-IDF, and WE with SBERT. Overall, the results suggest that TF-IDF and LDA are promising techniques for categorizing text data across different time periods, while WE with SBERT produced poor results for all three time periods.

8. Acknowledgements

I want to thank Libris for supplying the information used in this study. Moreover, I extend my thanks to my supervisors and colleagues for providing guidance, insights, and valuable comments.

9. References

[1] J. Frommeyer, Chronological Terms and Period Subdivisions in LCSH, RAMEAU, and RSWK, Library Resources & Technical Services, 48.3, (2013):199-212.

[2] M. J. Bates, D. N. Wilde, and S. Siegfried, An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report Number 1., The Library Quarterly 63.1 (1993): 1-39.

[3] T. Bogaard, L. Hollink, J. Wielemaker, J. van Ossenbruggen, and L. Hardman, Metadata categorization for identifying search patterns in a digital library, Journal of Documentation 75.2 (2019): 270-286.

[4] A. Dalli, 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, 2006, pp. 29--33, doi: 10.3115/1629235.1629238

[5] Metadatabyrån (Metadata Office), 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.

[6] A. Håkansson, R.L Hartung, Artificial Intelligence: concepts, areas, techniques and applications, 1st. ed., Studentlitteratur AB, Lund, 2020.

[7] J.Saarti, Fictional literature, classification and indexing, Knowledge Organization 46.4, (2019): 320-332.

[8] P. Rafferty, Epistemology, literary genre and knowledge organization systems, in: Actas del X Congreso de ISKO-España. Ferrol 20 de junio-1 de julio de 2011; 2013, pp.553-565.

[9] R. Maker, Finding what you're looking for: a reader-centred approach to the classification of adult fiction in public libraries. The Australian library journal 57.2 (2008) 169-177.

[10] A. Shenton, The role of 'Reactive classification' in relation to fiction collections in school libraries, New Review of Children's Literature and Librarianship 12.2 (2006) 127-146.

[11] P. Almeida, and G. Claudio, Fiction in a phenomenon-based classification, Cataloging & Classification Quarterly 59.5 (2021) 477-491.

[12] C. Manger, That Seems Made Up: Deep Learning Classifiers for Fiction & NonFiction Book Reviews, Masters' dissertation, Technological University Dublin, Republic of Ireland, 2018.

[13] M. Ströbel, E. Kerz, D. Wiechmann and Y, Qiao, 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, 2018, pp. 56-63.

[14] V. Kulkarni, Y. Tian, P. Dandiwala and S. Skiena, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, New Mexico, USA, 2018, pp. 202-212.

[15] B. Adams, Time, 1st ed,. Polity Press, Cambridge, UK, 2006.

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

[17] S. Bergsten, L. Elleström, Litteraturhistoriens grundbegrepp, 2nd. ed., Studentlitteratur AB, Lund, 2004.

[18] R. Shaw, Events and periods as concepts for organizing historical knowledge, 1st ed., University of California, Berkeley, 2010.

[19] T. Cook (Ed.), Quasi-experimental design, Wiley Encyclopedia of Management (2015) 1-2.

[20] Metadatabyrån (Metadata Agency), Lista kronologiska ämnesord särskilda länder, 24 Mar. 2021, URL:metadatabyran.kb.se/amnesord-och-genre-form/svenska-amnesord/typer-av-mnesord/kronologiska-amnesord/lista-kronologiska-amnesord-sarskilda-lander.

[21] A. Burkov, The hundred-page machine learning book, volume 1, Andriy Burkov, Quebec City, Canada, 2019.

[22] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3.1 (2003) 993-1022.

[23] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084, 2019. URL: https://doi.org/10.48550/arXiv.1908.10084.

[24] S.Shalev-Shwartz, S. Ben-David, Understanding machine learning: From theory to algorithms, 1st. ed., Cambridge University Press, 2014.