

# 18th-Century French Novel Analysis across Time, Gender and Sentiments

Team 5: Zselyke, Yara, Anass, Fabian, Vicky

## I. Project Overview:

### *Research Question*

How did the sentiments of French novel summaries evolve over the 18th century, and how did they differ according to the gender of the authors?

### *Thesis Statement*

We argue that the sentiments portrayed in the descriptions of 18th-century French novels shift significantly over time, with noticeable differences between male and female authors. These differences reflect the distinct societal roles and constraints females experienced, showing a gendered dimension in literary expression. A change in literary sentiment over time has been seen in 20th-century literature reflecting historical developments (Acerbi et al., 2013), and thus our team is interested to discover if the same can be observed in French novels over the 18th century.

Marlene LeGates displays a clear difference in how women and men write about womanhood with men being more traditional and women more progressive at the beginning of the 18th century (1976) which leads us to explore gender-based differences across various topics.

Our project will be analyzing plot summaries instead of the entire corpus, which according to Simone Rebora (2021), should yield similar sentiment results. This reduces the data volume and will speed up the analysis.

This research is timely and increasingly relevant today as new digital tools allow us to analyze and perform complex tasks on large amounts of data. With an increasing interest in gender representation, this paves the way for a more inclusive understanding of the past.

## II. Data Acquisition & Data Criticism:

The dataset published by the Trier Center for Digital Humanities (TCDH) at Trier University contains 200 digital texts of 18th Century French novels created or first published between 1751 and 1800. The two years with the most published books in this data set are 1798, with 10 books published, and 1800 with 11 books published. The dataset consists of 35 female, 159 male, and 6 uncategorized authors. The two authors with the most books in the collection are François-Thomas-Marie de Baculard d'Arnaud with 14 novels included and Voltaire with 13 Novels. The texts range from 601 to 309807 Words with an average of 60643.99 Words with 120 short, 45 medium and 35 long texts.

### *Source and Provenance of the Data*

The dataset was made as part of the Mining and Modeling Text project at Trier University between 2019 and 2023. To digitize late 18th-century French novels, OCR technology was combined with historical bibliographic records from Martin et al. (1977). Texts were being collected via Gallica (the digital platform of the Bibliothèque nationale de France) and other digital archives

A model (OCR4all) was specifically trained for 18 century French texts and to ensure that everything made sense quality wise, native French speakers made manual corrections of the texts. Additionally the TEI/XML format was used to encode the texts, guaranteeing that the dataset sticks to the digital text representation standards.

The metadata for the dataset was carefully arranged using a Python script designed to identify the gender of authors based on gender-specific titles and Wikidata was used to gather information regarding the authors. The dataset used Martin et al. (1977) for narrative forms and other bibliographic data, and human assessments of the texts were done to guarantee accuracy.

### *Data Preprocessing*

The OCR output was quality controlled in the preprocessing steps, and native speakers manually proofread the outputs to make corrections if deemed necessary. The dataset was then further structured and tweaked based on historical gender proportions, publication year, and narrative type, all of which were gathered from historical sources.

Additionally, complex SPARQL queries were made possible thanks to the Linked Open Data paradigm that was used to link these full text resources with the knowledge graph “MiMoTextbase.” Metadata from about 200 French novels (published between 1751 and 1800) are included in this knowledge graph, enabling analysis between the texts included in the dataset/ beyond just the 200 texts included in the dataset.

### *Data Criticism*

A critical assessment of the dataset reveals several considerations regarding its representativeness, selection criteria, and limitations:

The aim of collecting a representative sample of French novels written between 1751 and 1800 formed the basis of the selection process. As mentioned before, the year of publication, the author’s gender, and the narrative form served as the main selection criteria. Historical bibliographic information from Martin et al. (1977), which gave an overview of the literary production during this time period, served as the basis for these criteria. However, the dataset was meant to be balanced and it is thus far unclear how certain books were chosen from the larger range of literature, which might lead to biases in the selection process.

According to Jean Sgard’s “Dictionnaire des Journaux 1600-1789”. This comprehensive study covers literature, including novel production during the late 18th century, and suggests estimates of around 400 to 600 novels were published in France between 1751 and 1800. The dataset used in this research represents a subset of 200 novels, or about 50%/ 33% of the total published novels during this time. This could pose challenges in terms of the generalizability of the results.

Though certain biases may still exist, the dataset aims to be representative by simulating the historical distribution of books by gender, year, and form. Prominent male authors such as Voltaire and Baculard d’Arnaud, who between them have contributed 27 books to the dataset, accounting for more than 10% of the overall corpus, are heavily represented in the dataset. This might under-represent lesser-known works and writers by skewing the results towards specific topics or styles that are typical of these two more occurring authors.

As discussed earlier, a combination of automated and human procedures were used to gather the metadata. The writers’ gender was mostly determined using Wikidata, with further gender identification carried out by use of a Python script. Even though it works well, this approach could potentially contain errors, particularly for works that are anonymous or pseudonymous or when the gender is not made explicit in the title. Furthermore, human judgment and bibliographic metadata

were used to construct narrative forms, which introduces subjectivity into the novel categorization process and may have an impact on the accuracy of genre classifications.

Potential problems are introduced by the fact that our research focuses on assessing novel descriptions rather than the complete texts. The amount of information included in these descriptions can vary, with more well-known novels often receiving greater attention and in-depth analysis, which can result in unequal representation in the statistics. Furthermore, even if the descriptions are written in the historical language of the period, sarcasm, satire, and other complex sentiments that are characteristic of that time's literature can be difficult for modern sentiment analysis methods to capture. The dataset however offers a solid basis, considering the analysis's reach is restricted by its relatively small size.

Finally, the MiMoTextbase knowledge graph and the project's use of Linked Open Data offer chances to grow the dataset by adding more texts from Gallica or comparable digital archives. A more detailed statistical study and a more comprehensive investigation of literary trends throughout the French Enlightenment would be possible with such extension.

### III. Methodology:

| Workflow Section                | Task & Considerations  | Duration | Work Allocation              |
|---------------------------------|--|----------|------------------------------|
| Research Question (RQ) & Thesis | <u>(1) RQ Proposal:</u><br>Craft potential research questions and thesis statements for our project proposal submission. <ul style="list-style-type: none"> <li>Support the statements with relevant literature.</li> <li>Select the question that aligns the best with our team's goals.</li> </ul>   | Week 2-3 | All team members             |
|                                 | <u>(2a) RQ Refinement:</u><br>Brainstorm of additional backup RQs concurrently with (3). <ul style="list-style-type: none"> <li>Keyword search in the text instead of text summarisation and sentiment analysis?</li> <li>What are the additional research our team needs to undergo to support this RQ?</li> </ul>  | Week 4   | Zselyke, Yara, Anass, Fabian |
|                                 | <u>(2b) RQ Refinement:</u><br>Assessing research question feasibility in the context of our short project duration concurrently with (2a). <ul style="list-style-type: none"> <li>What libraries and models are required for the text processing and sentiment analysis?</li> <li>Can we strike a balance between computational efficiency and accuracy of the text processing steps?</li> </ul> | Week 4   | Vicky, Zselyke               |
| Dataset Processing              | <u>(3) Text Summarisation</u> <ul style="list-style-type: none"> <li>Which summarisation module can balance</li> </ul>   | Week 5-6 | Vicky                        |

|                                      |   |          |                      |
|--------------------------------------|---|----------|----------------------|
|                                      | <p>accuracy of the translations and speed of the translations?</p> <ul style="list-style-type: none"> <li>• How do we validate the summarisations?</li> </ul>   |          |                      |
|                                      | <p><u>(4) Sentiment Analysis</u></p> <ul style="list-style-type: none"> <li>• What sentiment analysis model can effectively predict the sentiment of French texts?</li> <li>• Is there an existing model trained on historical French?</li> </ul>   | Week 5-6 | Vicky                |
|                                      | <p><u>(5) Encoding</u></p> <ul style="list-style-type: none"> <li>• How do we prepare the data for the visualizations?</li> <li>• Which categorical columns do we need to encode?</li> </ul>  | Week 6   | Vicky                |
| Analysis & Findings                  | <p><u>(6) Data Visualisations</u></p> <ul style="list-style-type: none"> <li>• What charts and graphs are the most suitable for our RQ?</li> <li>• What statistical analysis can we do to further support our findings?</li> </ul>  | Week 6   | Fabian               |
|                                      | <p><u>(7) Data Insights</u></p> <ul style="list-style-type: none"> <li>• What are the findings from the visualizations?</li> <li>• What can we conclude from our analysis?</li> <li>• How does this relate to our thesis and research question?</li> </ul>  | Week 6   | Zselyke, Yara, Anass |
| Documentation, Report & Presentation | <p><u>(8) Documentation</u></p> <ul style="list-style-type: none"> <li>• Documentation of workflow at each step using Github and Google Drive.</li> <li>• Distribute workload according to the skills of team members.</li> </ul>   | Week 2-7 | All team members     |
|                                      | <p><u>(8) Team Meetings</u></p> <ul style="list-style-type: none"> <li>• Align project goals and distribute workload.</li> <li>• Weekly meetings on Wednesdays at 8pm</li> </ul>  | Week 7   |                      |
|                                      | <p><u>(9) Report Writing &amp; Presentation</u></p> <ul style="list-style-type: none"> <li>• Documentation of workflow at each step using Github and Google Drive.</li> <li>• Distribute workload according to the skills of team members, as shown below:</li> </ul> <p>Yara: Ethical considerations, documentation and sustainability</p> | Week 7   |                      |

|  |  |  |  |
|--|--|--|--|
|  | Zselyke: Ethical considerations, challenges and solutions<br>Anass: Data acquisition & data criticism, challenges and solutions<br>Fabian: Data visualizations, data results<br>Vicky: Methodology, workflow steps |  |  |
|--|--|--|--|

#### IV. Workflow Steps:

##### 1. Research Question Refinement & Considerations

Before we embarked on the workflow, we assessed its feasibility by doing some preliminary research on potential models we can use to carry out the text summarization and sentiment analysis. We were aware of the challenge that our dataset was in French, with some texts in historical French, and were unsure if there are pre-trained models available to carry out the text processing tasks. Therefore, our team decided to assess its feasibility whilst brainstorming additional questions should we find that there are no suitable models for our text processing tasks.

For the feasibility assessment, we looked at open-source models such as BERT, and explored fine-tuned models such as camemBERT on HuggingFace, an open-source site for deep learning models. Concurrently with this assessment, the rest of the team brainstormed additional ideas and assessed its feasibility as well:

| Idea  | Pros   | Cons  |
|---|--|---|
| Observing the sentiment trend in <b>novel titles</b> across gender and time                         | Does not require a text summarisation module which has its limitations and inaccuracies, and reducing technical complexities | Our team agreed that titles are an oversimplification of the novel content to depict the true sentiments of the novel. We believed that it captures less nuances than a novel summary and were also unable to find supporting literature for this question. |
| Observing the <b>frequency of the word 'love'</b> in the entire text across gender and time         | Interesting idea with less complexities and a more straightforward methodology   | The theme of love might not be captured with a keyword search, posing the question of is the word 'love' utilized in the context of love. The literary complexity, as well as cultural differences in 'love' were a deterrent against this question.        |
| Observing the <b>number of female and male characters</b> in the entire text across gender and time | Interesting idea that compares the usage of female and male characters in the novel.   | Greater technical complexities will be introduced in the gender prediction of the characters as well as the character name retrieval. Uncertainties and ambiguity in names might result in inaccuracies and affect results as well.                         |

Ultimately, after these assessments, we were able to find models fine-tuned on French datasets with a relatively high accuracy of 78% and greater. This would mean that our initial research question and thesis would be feasible, and our group decided to move ahead with it. All of our

additional questions brainstormed had its own set of limitations, uncertainties and ambiguity involved, as well as certain advantages over our initial research question. However, the greatest deterrent against the additional questions was that we were unable to find relevant literature to support these questions, while we had already found supporting literature for our research question and thesis. Furthermore, as our research question and thesis are already supported by literature and research, its greatest limitation is the technical complexities, which can be overcome if the project duration was longer and if we were able to fine-tune a model from scratch. Our initial question also aligns with our team's individual interests and goals the best, and therefore, we proceeded on with our original research question and thesis.

## 2. Exploratory Data Analysis

| Tool          | Description & Purpose   |
|---------------|---|
| Python/Pandas | Visualize and calculate statistics regarding the data, plot graphs to aid in the understanding of dataset |
| Excel         | View individual data rows and points to understand the data more distinctly and specifically              |

## 3. Text Summarisation

| Library  | Description & Purpose   |
|--|---|
| bert-extractive-summarizer   | This model uses BERT to select key sentences from a document for extractive summarization, retaining the core meaning of the text.                          |
| mrm8488/camembert2camembert_shared-finetuned-french-summarization<br><a href="https://huggingface.co/mrm8488/camembert2camembert_shared-finetuned-french-summarization">https://huggingface.co/mrm8488/camembert2camembert_shared-finetuned-french-summarization</a> | A CamemBERT model from HuggingFace fine-tuned on a multilingual dataset with online newspapers to generate concise and accurate summaries for French texts. |

In our final selection of model for the novel summarisation, we selected the *mrm8488/camembert2camembert\_shared-finetuned-french-summarization* model as it was able to efficiently summarize each text in 5 seconds as compared to 3 minutes for *bert-extractive-summarizer*. Furthermore, CamemBERT boasts greater accuracy as it is more specific due to being fine tuned on a multilingual dataset, including French texts.

```
novel-analysis > summary > 📄 Arnaud_Sidney.txt
1  A la bonne heure qu'Isaac Newton soit un grand-homme, et que notre île
2  s'applaudisse de lui, je souscris de tout mon cœur à un si juste éloge.
```

Figure 1: Summarisation of the novel 'Arnaud Sidney'

## 4. Sentiment Analysis

| Library                          | Description & Purpose                       |
|----------------------------------|---|
| ac0hik/Sentiment_Analysis_French | This model from HuggingFace is a fine-tuned |

|   |   |
|---|---|
| <a href="https://huggingface.co/ac0hik/Sentiment_Analysis_French">https://huggingface.co/ac0hik/Sentiment_Analysis_French</a>   | version of CamemBERT model, performing a sentiment prediction and classifying French text as positive, negative, or neutral.  |
| cmarkea/distilcamembert-base-sentiment<br><a href="https://huggingface.co/cmarkea/distilcamembert-base-sentiment">https://huggingface.co/cmarkea/distilcamembert-base-sentiment</a> | A distilled version of CamemBERT from HuggingFace , optimized for fast and efficient sentiment analysis in French texts. Output labels are 1-5, with 1 being the most negative and being the most positive. |

Both models are fine tuned CamemBERT models on French datasets, with the key difference namely being the output of the datasets. However, upon performing the classification and analyzing the predictive strength of the models, *ac0hik/Sentiment\_Analysis\_French* produced greater strength in the prediction of sentiment labels as compared to *cmarkea/distilcamembert-base-sentiment* and therefore, we went ahead with this model.

Average probability strength of classification for *ac0hik/Sentiment\_Analysis\_French*: 0.7853  
Average probability strength of classification for *cmarkea/distilcamembert-base-sentiment*: 0.3864

Figure 2: Strength of classification of sentiment labels for each model


After retrieving the sentiment labels, we extracted the probabilities for each sentiment predicted by the model. A greater probability value indicates a stronger sentiment observed for that text summary. These probabilities will be utilized in data visualizations to observe the strength of sentiments over time.

| summary   | sentiment | probability |
|---|-----------|-------------|
| La chronique de Roger-Pol Droit, à propos de "... | neutral   | 0.871761    |
| A la veille de la présidentielle, "Le Monde" d... | positive  | 0.858749    |
| L'E fils du potentat comme celui du savetier S... | neutral   | 0.841893    |
| Dans sa chronique pour le cahier "Sport&Forme"... | neutral   | 0.706780    |

Figure 3: Dataset snippet with the addition of 'summary', 'sentiment' and 'probability' columns

## 5. Data Pre-processing

| Tool          | Description & Purpose   |
|---------------|---|
| Python/Pandas | Creation of dummy variables and encoding to turn categorical data into numerical data |

|   |   |
|---|---|
|   |  <p>Figure 4: Dataset snippet with the addition of 'gender_label' and 'sentiment_label' columns</p> |
| R | Creation of dataset subsets where a filter for gender or sentiment is utilized to segment the data for the visualizations   |

## 6. Data Visualizations

| Chart Type        | Description & Purpose  |
|-------------------|--|
| Pie Chart (in R)  | Visualizing the proportion of sentimentality in our dataset and for each gender for comparison and contrasting |
| Line Chart (in R) | Visualizing the strength of the sentiments across time for each gender to spot trends                          |

## V. Challenges and Solutions:

### *Dataset Complexity and RQ Feasibility*

When we began with the analysis we came to the realization that our original research question might not be the most practical and feasible due to limited time and resources such as computational power or natural language processing (NLP) models. This was due to the difficulty in handling the dataset due to it being in French with some even in historical French and a large data volume for each novel. Historical French NLP models have little advancement in the pre-trained models field yet, therefore we are likely to face inaccuracies in our results.

Vicky suggested that we would look into other possible research topics while she would attempt to research different methods for sentiment analysis, as she has much more experience and knowledge in that field compared to the rest of us. Thus, our team decided to work in parallel by assessing the feasibility of our RQ with supporting research on possible NLP models, while brainstorming potential other RQs. While the rest of us were working on alternative RQs we considered whether it would be possible to modify the original RQ to solely focus on the provided metadata rather than on sentiment analysis. This included utilizing metadata to examine patterns in publication years, author gender, and narrative forms. However, during the brainstorming session, we realized that every question has its own set of limitations and that it was difficult to find relevant supporting literature to support our question and thesis. We also realized that only using the metadata would extremely simplify our research project and limit our exploration of the novels as the dataset was primarily qualitative rather than quantitative.



Fortunately, Vicky successfully found some models of a decent accuracy score of above 78% that would make our original RQ feasible. She identified the CamemBERT model (from Hugging Face) for the text summarization. The model proved to be effective and efficient at generating summaries of the novels, which cut down a lot of the processing time it would have taken with other models. For sentiment analysis Vicky found and selected the *ac0hik/Sentiment\_Analysis\_French*, another fine-tuned CamemBERT model from HuggingFace, which provided us with sentiment classifications (positive, negative or neutral) for the summarized texts. This allowed us to eventually achieve our objectives in our methodology and workflow process.

#### *Familiarity with the methods and tools used for text sentiment analysis*

Another challenge was that some team members were not familiar with the methodologies required for sentiment analysis and text summarization due to our group having people with diverse fields of study. While some team members had strong backgrounds in literature and humanities, others specialized in areas like computer science or data analysis. This meant that not everyone knew the specific tools used or had the experience for sentiment analysis.

To address this we made sure to document the steps of the methodology, explaining which tools were used and in what way they were used. By organizing weekly group sessions we made sure that everyone would be up to date as to what had been done and what had to be done next. This way we made sure that everyone, regardless of their study background, knew which steps were being taken in the methodology.

#### **VI. Ethical Considerations:**

When it comes to ethical considerations, given that the dataset consists of novels written by authors who lived in the 18th century, modern privacy laws such as the General Data Protection Regulation (GDPR) do not apply. Therefore, there are minimal concerns related to data privacy or the protection of personal information. However, some ethical considerations still arise when using this dataset. For instance:

- The dataset represents male authors (159 male vs. 35 female), which may lead to gender bias in the analysis.
- The novels reflect societal views from the 18th century, which may include outdated or offensive ideas regarding gender, class, and race, which makes it essential to approach this content with sensitivity and critical analysis as much as possible.

#### **VII. Results**

To answer our research question we performed a multinomial regression analysis with the sentiment labels as the dependent variable and gender and year as independent variables. As gender has three levels with the inclusion of the unidentified label we need to conduct the research on reference level 1, this means that the probabilities of the unidentified and Female label are being compared to the male label. This gave the results displayed in *Table 1*.

The first regression is between neutral and negative. It shows that unidentified and female authors were less likely to produce neutral rather than negative sentiments than male authors in this dataset. It also showed this effect was weaker in female authors. When the time variable is integrated it can be shown that texts become progressively more neutral throughout the years.

The second regression is between positive and negative sentiments. While unidentified authors were less likely than male authors to produce positive sentiments, this effect is smaller than in

the first regression. Female authors were more likely to write novels with positive sentiment than negative when compared to male authors. This effect also gets stronger over time.

To summarize our findings we can say that, in this dataset, Male authors are more likely than Unidentified and Female authors to produce neutral sentiment novels compared to negative ones.

Female authors are more likely than Male authors to produce positive sentiment novels compared to negative ones. As time progresses, the likelihood of both neutral and positive sentiments increases compared to negative sentiment for all genders.

### Coefficients:

|   | (Intercept) | gen_labMale | gen_labFemale | date       |
|---|-------------|-------------|---------------|------------|
| 1 | -17.35307   | -0.82717429 | -0.728980     | 0.01053994 |
| 2 | -11.58090   | -0.09948237 | 1.015762      | 0.00651361 |

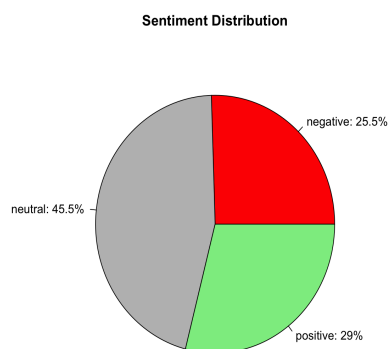
### Std. Errors:

|   | (Intercept)  | gen_labMale | gen_labFemale | date         |
|---|--------------|-------------|---------------|--------------|
| 1 | 0.0007374026 | 0.2788692   | 0.2447338     | 0.0001530855 |
| 2 | 0.0005064664 | 0.2596782   | 0.2565923     | 0.0001446472 |

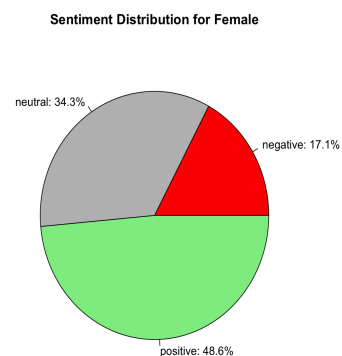
Residual Deviance: 413.7268

AIC: 429.7268

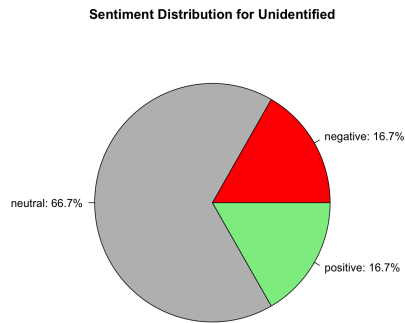
Table 1: Multinomial Regression conducted in R Studio



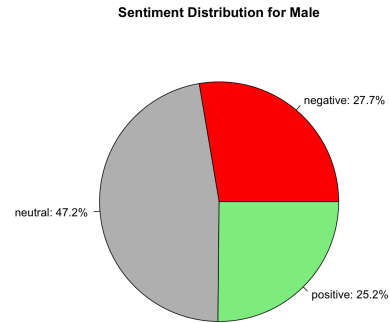
Graph 1: Sentiment distribution unfiltered



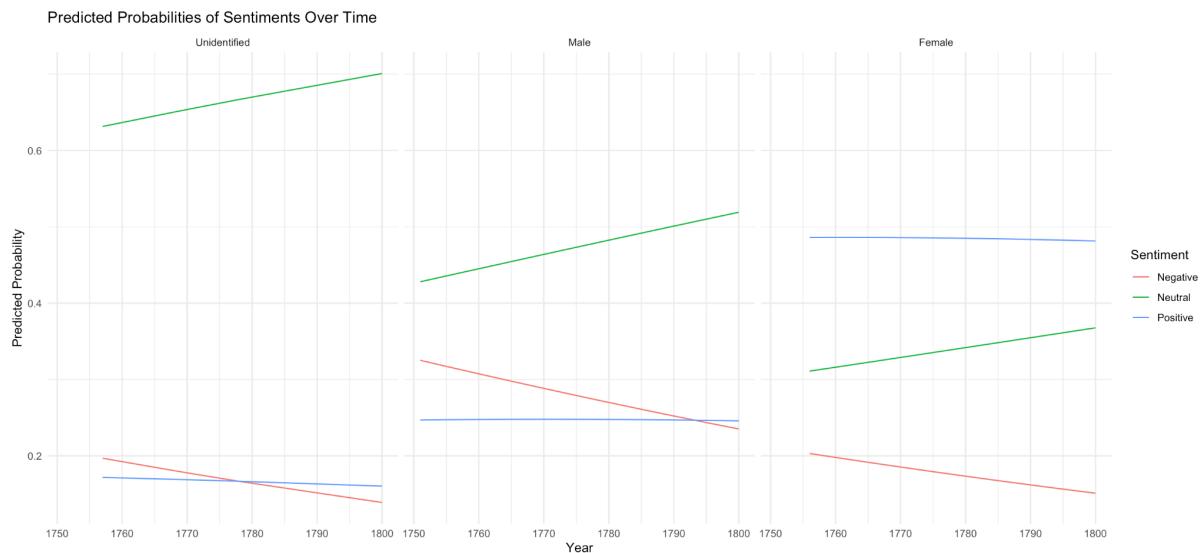
Graph 2: Sentiment distribution filtered for 'female' authors



Graph 3: Sentiment distribution filtered for 'Unidentified' authors



Graph 4: Sentiment distribution filtered for 'Male' authors



Graph 5: Strength of sentiments (predicted probabilities) over time filtered by gender

## II. Limitations and Conclusion

Our analysis faced several key limitations. Firstly, the social constraints imposed on female authors during the 18th century likely influenced their choice of topics and writing styles. This makes it challenging to determine whether the differences in sentiment were truly gender-based expressions or if they resulted from these external pressures. Additionally, the evolution of the French language from the 18th century to today—especially in terms of grammar and spelling—posed significant difficulties when applying modern NLP tools. This may have affected the accuracy of our sentiment analysis. Furthermore, our study was limited to analyzing plot summaries rather than full texts. While this method allowed for quicker processing, it likely oversimplified the sentiment, missing the more nuanced expressions found in the full narratives.

In conclusion, our study has highlighted the gendered dimension of literary expression in 18th-century French novels. In relation to our RQ and thesis, we observed a notable shift in sentiment across the century, with male and female authors expressing emotions differently—likely influenced by societal constraints. These findings suggest avenues for future research, particularly with larger

datasets and more advanced tools. Our findings can also be further explored in a more nuanced research regarding gender and gender roles and its sentiments during this time period. Ultimately, our project underscores the importance of both sentiment trends and gender analysis in understanding the literature of this era, and leaves the door open for subsequent research on gender differences during this era.

## IX. Documentation and Sustainability

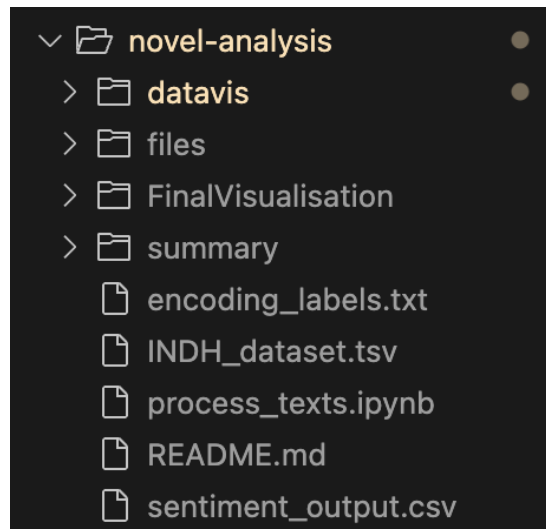


Figure 5: Snippet of Github repository file structure and contents

The link to our Github repository:

<https://github.com/vickyqu/introduction-to-digital-humanities-team-5>

| Folder/File                   | Description   |
|-------------------------------|---|
| <i>datavis</i>                | The first iteration of our data visualizations with R, however, there are still some missing labels and improvements that can be made to better visualize the sentiment trends over time.   |
| <i>files</i>                  | Dataset downloaded from <a href="https://github.com/MiMoText/roman18">https://github.com/MiMoText/roman18</a> containing the original full text from each of the 200 French novels.   |
| <i>FinalVisualisation</i>     | After refinement and discussions with the team, a final set of data visualizations were finalized, including the following: <ul style="list-style-type: none"> <li>- Multinomial regression analysis</li> <li>- Line charts depicting sentimentality strength over time and across genders</li> <li>- Pie charts depicting proportion of sentiments across genders</li> </ul> |
| <i>FinalVisualisation/NEW</i> | Final visualizations of the line charts depicting sentimentality strength over time and across genders.   |
| <i>summary</i>                | This is a folder containing the summarized contents for each novel in the dataset.  |
| <i>encoding_labels.txt</i>    | Documentation for how the categorical variables (gender, sentiment) were encoded.   |

|                             |  |
|-----------------------------|--|
| <i>INDH_dataset.tsv</i>     | Original dataset of the metadata from the 200 French novels.                               |
| <i>process_texts.ipynb</i>  | Python file with the text summarisation, sentiment analysis and encoding processing tasks. |
| <i>sentiment_output.csv</i> | Final processed dataset with the summary, sentiment labels and encoded values.             |

## X. Reflection

Our project workflow worked well overall, especially with clear task assignments and regular team meetings. We faced some difficulties at first, like choosing the right models and dealing with dataset limits. Still, we adjusted by narrowing our focus and using the best tools for text summarization and sentiment analysis.

One of the most important lessons we learned is the need to remain flexible with our research question and methodology. Initially, we had doubts about the feasibility of conducting sentiment analysis on such a limited and historically distant dataset. However, by testing different models and discussing alternative approaches, we were able to adapt and refine our workflow.

Another thing we learned is the importance of clear documentation. Given the diverse backgrounds of our team members, documentation was essential. By clearly outlining each step of the process—especially for technical tasks like model selection and data preprocessing—team members without a technical background were able to follow and contribute to the project. This also ensured that any future researchers who might use our work will have an easier time understanding our methodology.

Lastly, for future improvements, we suggest expanding the dataset, fine-tuning sentiment models for historical texts, and exploring additional textual features. This would enhance the analysis and lead to more comprehensive insights into 18th-century French novels.

## XI. References

- Acerbi, A., Lamos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLoS ONE*, 8(3), e59030. <https://doi.org/10.1371/journal.pone.0059030>
- Brodeur, P.-O. (2013). *Le roman édifiant aux XVIIe et XVIIIe siècles* [Doctoral dissertation, Université de la Sorbonne nouvelle-Paris III & Université de Montréal]. TEL Archives. <https://tel.archives-ouvertes.fr/tel-00950419>
- LeGates, M. (1976). The cult of womanhood in Eighteenth-Century thought. *Eighteenth-Century Studies*, 10(1), 21. <https://doi.org/10.2307/2737815>
- Martin, P., Coulet, H., Rétat, P., & Sgard, J. (1977). *Roman et société en France aux XVIIe et XVIIIe siècles*. Droz. <https://doi.org/10.3917/puf.monta.1999.02.0523>
- Palmer, M. (1992). *Dictionnaire des journaux 1600-1789* (Sous la direction de Jean Sgard) [Compte-rendu]. *Réseaux. Communication - Technologie - Société*, 10(53), 123–126. [https://www.persee.fr/doc/reso\\_0751-7971\\_1992\\_num\\_10\\_53\\_1986](https://www.persee.fr/doc/reso_0751-7971_1992_num_10_53_1986)

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M. M., Lauer, G., Lendvai, P., Messerli, T. C., & Sorrentino, P. (2021). Digital humanities and digital social reading. *Digital Scholarship in the Humanities*, 36(Supplement\_2), ii230–ii250.  
<https://doi.org/10.1093/llc/fqab020>

Röttgermann, J. (2024). The collection of eighteenth-century French novels 1751–1800. *Journal of Open Humanities Data*, 10(1), 31. <https://doi.org/10.5334/johd.201>.

Trier Center for Digital Humanities. (2019-2023). *Mining and modeling text project*. Trier University.  
<https://tcdh.uni-trier.de/en/newsbeitrag/project-closure-mining-and-modeling-text-2019-2023>

Urologin, S. (2018). Sentiment analysis, visualization and classification of summarized news articles: A novel approach. *International Journal of Advanced Computer Science and Applications*, 9(8).  
<https://doi.org/10.14569/ijacsa.2018.090878>