HERMES Data Challenge 2024: Research Question

Boro Sofranac[1], Daniel Thürck , Lilian Do Khac , and Marc Fabian Mezger

[1] *Podgorica Tech*
(alphabetically ordered)

November 18, 2024

boro@pgtech.me, daniel@hpclabs.de, Dokhac@students.uni-marburg.de, marc.mezger@gmail.com

Abstract

In the following, our research question for the HERMES Data Challenge 2024 is described.
Corresponding GitHub repository: `https://github.com/dthuerck/daLLMayr`

1 Contextualisation of the research question

Newspaper content analysis provides an invaluable resource for historical research, offering a window into societal transformations. By analyzing these sources, researchers can identify the influence of external factors and track the emergence of issues like environmentalism and shifting social norms. Event-Structure Analysis (ESA) is a powerful technique, aiding in the interpretation of complex historical narratives and causal relationships. [Griffin and Korstad(1998)]

However, ESA's application to extensive datasets, a consequence of digitization, poses challenges. To ensure accuracy, rigorous data preprocessing is essential. Our research project proposes a flexible data pipeline, designed to handle big data efficiently, providing a robust foundation for our study.

Focusing on the period between 1919 and 1945, a time of rapid technological and sociopolitical change, we aim to understand how newspapers reflected these advancements and their societal impact. Our research will explore public perceptions of technology, considering its dual nature as a potential economic solution and a threat to traditional ways of life. We will also investigate the influence of urbanization and its impact on gender roles, particularly women's role in the workforce.

Following research questions guide this analysis:

1. **Sentiment about Technological Developments:** How did German newspapers between 1919 and 1945 reflect positive and negative sentiments about technological advancements, and what factors influenced these perceptions?
2. **Urban vs. Rural Perceptions of Technological Advancements in Germany:** How did perceptions of technological advancements differ between urban and rural areas in Germany, as reflected in newspaper coverage from 1919 to 1945?
3. **Technology and Gender Roles in Germany: Media Representation:** How did German newspapers from 1919 to 1945 reflect the impact of technological advancements on gender roles, particularly regarding women's participation in the workforce?

2 Data processing workflow

We have developed a multi-step process that leverages the latest advancements in natural language processing techniques. Our approach aims to bridge the gap between historical and modern German texts, making them accessible to a wider audience. By utilizing Large Language Models (LLMs), we first translate old German texts into modern German, ensuring comprehension and broadening the reach of these historical documents.

The next step involves employing LLMs to summarize newspaper articles, extracting key information and creating concise overviews. These summaries not only benefit readers by providing a quick glimpse into the articles' content but also serve as a foundation for further analysis. By condensing the articles into their core points, we enable researchers and interested individuals to quickly identify relevant topics and gain an overview of the material. To further process and analyze the extracted information, we transform the summaries into numerical embeddings. These embeddings allow us to represent the content in a mathematical space, enabling us to visualize patterns and connections between the articles. By mapping the articles onto this space, we can explore complex relationships and build a network of information.

To enhance our understanding of the articles, we integrate sentiment data into the embeddings. By considering the emotional tone of the articles, we gain insight into the authors' sentiments and intentions. This additional dimension allows us to analyze the content not only on a factual level but also to consider the feelings and opinions expressed within the texts, providing a more comprehensive understanding of the historical and current events.

A central aspect of our process is the clustering method we employ to identify thematic clusters. By grouping similar topics together, we can explore the diversity of content and understand how different themes are interconnected. This clustering enables us to visualize a complex network of information and trace the development of topics over time. Applying research questions to these thematic clusters is a powerful tool for gaining new insights. By targeting specific clusters, we can answer focused research questions and develop a deeper understanding of the content. This process allows us to assess the relevance and significance of the extracted information, providing valuable insights for historians, journalists, and other interested parties.

Analyzing newspaper archives presents a challenging task, particularly due to the vast amount of data these archives contain. A critical aspect often overlooked is the extraction of individual articles from newspaper pages. While modern language models excel at translation and summarization, they struggle to differentiate between multiple articles on a single page. This challenge can lead to mixed topics and inaccurate analyses, compromising the accuracy and reliability of the results. To address this challenge, it is crucial to enhance optical character recognition (OCR) technology. By employing more precise OCR techniques, we can accurately identify the boundaries between articles, resulting in improved extraction. Enhanced OCR not only increases data accuracy but also improves the efficiency of the entire process. By precisely recognizing article boundaries, we can ensure that each piece of information is correctly attributed, laying the foundation for reliable analysis.

3 Limitations

The combination of advanced OCR and specialized language models has the potential to revolutionize data extraction from newspaper archives. This synergy enables us to conduct mass analyses, laying the groundwork for comprehensive exploration of the data. Preprocessing large amounts of data allows scientists to elevate their research to new heights and explore diverse research questions. By combining AI techniques and language models, we can transcend the boundaries of traditional research and gain new perspectives. However, the vast data volumes contained in newspaper archives present a challenge. It was difficult to process a representative sample with the available computing resources, leading to a certain degree of caution when interpreting the interim results. The limitations arising from this process are multifaceted and require

careful consideration.

Firstly, Large Language Models (LLMs) can lead to hallucinations, where false or non-existent information is generated. These hallucinations can impact the accuracy and reliability of the generated summaries and analyses. To mitigate this risk, risk-minimizing procedures should accompany the data processing step. By implementing control mechanisms, we can ensure that the generated content is fact-based and not tainted by false information.

Secondly, selecting the optimal number of clusters is a challenge. Too few clusters may result in inaccurate outcomes, while too many clusters can lead to overlaps and confusion. Identifying the ideal number of clusters is a critical aspect, as it influences the accuracy and clarity of the thematic grouping. Through careful analysis and evaluation of the data, we can determine the optimal cluster count and achieve precise thematic grouping.

Thirdly, the analysis process may not capture all the nuances and sub-themes within the clusters. A deeper breakdown and analysis of sub-clusters is necessary to achieve a more comprehensive coverage of topics. By exploring these sub-themes, we can gain a more detailed understanding of the complex relationships and subtleties within the data. A detailed analysis allows us to explore the intricacies of the themes and paint a more comprehensive picture.

Fourthly, validation and evaluation of the interim results are of utmost importance. To ensure the accuracy and reliability of the outcomes, a comprehensive validation of the process is necessary. An evaluation that includes reference data and comparison possibilities is essential to assess the robustness and accuracy of our methods. Through this validation, we can ensure that our analyses are based on solid foundations and provide reliable results.

Lastly, computational performance and scalability pose a challenge, especially when processing extensive newspaper archives. The scalability of the process is crucial to enable efficient and rapid analysis. By utilizing powerful computing resources, we can reduce processing time and analyze larger datasets in a shorter period. Investing in high-performance computing capabilities is therefore an important step to enhance the efficiency of our process and enable the analysis of vast amounts of data.

4 Exemplary results (in German)

Due to time constraints and computing resource issues, we can only deliver exemplary results for Research Question 1, which should be consumed carefully since our work definitely lacks evaluation and control of the outputs.

4.1 Exemplary result (in German) for positive reflection

Zwischen 1914-1919 und 1938-1945 verändert sich die Darstellung und Wahrnehmung technologischer Fortschritte in deutschen Zeitungen deutlich. Hier eine Zusammenfassung der wichtigsten Unterschiede:

1914-1919 (Erster Weltkrieg und unmittelbare Nachkriegszeit):

* **Indirekte positive Darstellung:** Technologie wird nicht direkt gepriesen, sondern im Kontext von positiven Folgen wie internationale Kooperation, wirtschaftliche Erholung und nationale Einheit. Der Fokus liegt auf den *Ergebnissen* des technologischen Fortschritts, nicht auf der Technologie selbst. * **Diplomatie und Wirtschaft im Vordergrund:** Die Artikel betonen die Rolle von Technologien wie Telegraf, Zug und Dampfschiff bei der Ermöglichung von Diplomatie, Handel und internationalen Verträgen. * **Propaganda und Moralerhaltung:** Die positive Darstellung dient der Moralerhaltung und der Rechtfertigung der

Kriegsanstrengungen. Negative Aspekte der Technologie, insbesondere ihre zerstörerische Kraft im Krieg, werden heruntergespielt. * **Implizite Anerkennung wissenschaftlichen Fortschritts:** In einigen Artikeln wird die Bedeutung wissenschaftlicher Forschung und Entwicklung angedeutet, jedoch ohne detaillierte technische Beschreibungen.

1938-1945 (Zeit des Nationalsozialismus und Zweiter Weltkrieg):

* **Direkte und überschwängliche positive Darstellung:** Technologischer Fortschritt, insbesondere in Bereichen wie Luftfahrt, Automobilbau und Kommunikation, wird explizit gepriesen und als Beweis deutscher Überlegenheit dargestellt. * **Nationalismus und Propaganda:** Die Artikel sind stark nationalistisch geprägt und dienen der Propaganda des NS-Regimes. Technologische Errungenschaften werden als Zeichen nationaler Stärke und Größe inszeniert. * **Fokus auf Rekorde und Siege:** Die Berichterstattung konzentriert sich auf Rekorde, Siege in Rennen und sportliche Triumphe, die die deutsche technische Leistungsfähigkeit demonstrieren sollen. * **Verherrlichung militärischer Technologie:** Im Kontext des Zweiten Weltkriegs wird die militärische Technologie, insbesondere die Luftwaffe und U-Boote, glorifiziert. * **Ausblendung negativer Aspekte und der Konkurrenz:** Die Leistungen anderer Nationen werden heruntergespielt, und negative Aspekte oder Herausforderungen der deutschen Technologie werden ausgeblendet. * **Infrastruktur und Autarkie:** Technologischer Fortschritt wird mit der Stärkung der Infrastruktur und dem Streben nach Autarkie verknüpft.

Zusammenfassend lässt sich sagen:

Die Wahrnehmung technologischen Fortschritts wandelt sich von einer indirekten, kontextbezogenen und propagandistisch geprägten Darstellung im Ersten Weltkrieg zu einer direkten, überschwänglichen und nationalistisch aufgeladenen Verherrlichung im Nationalsozialismus. Der Fokus verschiebt sich von den positiven Folgen der Technologie auf die Technologie selbst, insbesondere im militärischen Bereich. Die Berichterstattung wird zunehmend propagandistisch und dient der Stärkung des NS-Regimes.

4.2 Exemplary result (in German) for negative reflection

Die größte Veränderung zwischen 1914-1919 und 1938-1945 ist der **Wechsel vom Ersten zum Zweiten Weltkrieg und der damit einhergehende Aufstieg des Nationalsozialismus in Deutschland**. Diese Veränderung prägt die Sichtweise auf technologischen Fortschritt in den Zeitungsartikeln fundamental.

1914-1919:

* **Desillusionierung durch den Krieg:** Die Erfahrung des Ersten Weltkriegs, insbesondere die verheerenden Folgen neuer Waffentechnologien wie Giftgas und Maschinengewehre, führten zu einer Ernüchterung über den technologischen Fortschritt. Technologie wurde mit Leid, Zerstörung und dem Scheitern der Kriegsziele assoziiert. * **Soziale und wirtschaftliche Folgen:** Die Kriegswirtschaft, Ressourcenknappheit und soziale Umwälzungen wurden ebenfalls mit Technologie in Verbindung gebracht, allerdings weniger direkt. Es gab Ängste vor den Folgen der Industrialisierung und der Veränderung der Arbeitswelt. * **Propaganda und Medien:** Die zunehmende Bedeutung von Massenmedien und Propaganda wurde als zweischneidiges Schwert wahrgenommen. Technologie ermöglichte die Verbreitung von Informationen, aber auch deren Manipulation.

1938-1945:

* **Instrumentalisierung der Technologie:** Unter dem Nationalsozialismus wurde Technologie gezielt für die Kriegsführung und die Durchsetzung der Ideologie instrumentalisiert. Technologischer Fortschritt wurde vor allem im militärischen Bereich vorangetrieben, während zivile Innovationen in den Hintergrund traten. * **Kontrolle und Zensur:** Die Nazis kontrollierten die Medien und unterdrückten kritische Stimmen. Technologien der Kommunikation wurden für Propaganda und die Verbreitung der NS-Ideologie genutzt.

* **Wirtschaftliche Regulierung:** Die Wirtschaft wurde streng reguliert und auf den Krieg ausgerichtet. Technologischer Fortschritt wurde durch Ressourcenknappheit und die Fokussierung auf Rüstungsproduktion gehemmt. * **Feindbild und Propaganda:** Technologische Überlegenheit der Alliierten wurde heruntergespielt oder als barbarisch dargestellt. Deutsche "Wunderwaffen" wurden propagandistisch hervorgehoben, um die Moral zu stärken.

Zusammenfassend lässt sich sagen:

Während im Ersten Weltkrieg eine generelle Desillusionierung über den technologischen Fortschritt herrschte, der mit Zerstörung und Leid verbunden wurde, wurde Technologie im Zweiten Weltkrieg unter den Nazis gezielt instrumentalisiert und kontrolliert. Die Propaganda nutzte die Technologie zur Verbreitung der Ideologie und zur Manipulation der öffentlichen Meinung. Die negative Wahrnehmung von Technologie konzentrierte sich dabei vor allem auf die technologische Überlegenheit der Feinde und die damit verbundene Bedrohung. Die Kontrolle und Lenkung der Technologie durch das NS-Regime hemmte zudem den zivilen technologischen Fortschritt.

References

[Griffin and Korstad(1998)] L. J. Griffin, R. R. Korstad, Historical inference and event-structure analysis, International Review of Social History 43 (1998) 145–165.