Analyzing the Coverage of Banking Subjects in Swiss Local Newspapers using different NLP Techniques

Hosseini Sayyed Ahmad, Pfister Kilian, Zuber Marc

Abstract

This paper aims to investigate the coverage and prominence of banking subjects in Swiss local newspapers over time using Natural Language Processing (NLP) techniques. The research question addresses the potential differences between German-speaking and French-speaking regions in Switzerland. A dataset of 900,000 articles from five German and four French local newspapers was analyzed using topic modeling with BertTopic, a transformer model from Huggingface. The findings reveal the frequency and prominence of banking topics in both regions and offer insights into potential disparities. This study contributes to understanding the evolution of banking coverage in Swiss local newspapers and sheds light on regional variations in media attention.

1 Introduction

The role of local newspapers in shaping public discourse and disseminating information is crucial for understanding societal dynamics and public opinion. In the context of the banking industry, media coverage plays a significant role in shaping public perceptions, influencing financial decisions, and even impacting the stability of the banking sector itself. Analyzing the coverage and prominence of banking subjects in local newspapers can provide insights into how the industry is represented in the media, identify potential biases, and uncover regional differences in media attention.

This study focuses on the Swiss context, specifically investigating the coverage of banking subjects in German-speaking and French-speaking local newspapers. Switzerland, renowned for its robust banking sector and multilingual regions, provides an intriguing setting to examine potential variations in media coverage. By leveraging NLP techniques and employing topic modeling, we aim to uncover temporal trends, highlight prominent banking topics, and explore potential discrepancies between the German-speaking and French-speaking regions.

2 Related Work

In 2022, Grootendorst introduced BERTopic (Grootendorst, 2022) as a topic model for extracting topics from text collections. Multiple studies have demonstrated that BERTopic can outperform other topic modeling methods, such as LDA (Blei et al., 2003) or top2vec(Angelov, 2020), particularly due to its ability to capture semantic relationships (Chen et al., 2023) (Ogunleye et al., 2023). Hence, we employ BERTopic to process our documents in order to achieve optimal results.

Topic modeling is a widely adopted approach for analyzing large volumes of text, including the focus of our study. This paper primarily addresses studies that have utilized BERTopic for topic extraction. However, no previous studies have specifically examined banking issues in Swiss newspapers. (Balaneji and Maringer, 2022) study concentrated on stock markets and employed LDA for topic modeling, whereas our research investigates a comprehensive time-period corpus sourced from various Swiss newspapers in French and German languages.

3 Dataset

Our dataset consists of 900,000 articles collected from nine local newspapers, five of which are German (NZZ, BZ, BU, BAZ, TA) and four are French (TLM, TDG, HEU, TPS). Each newspaper contributed 100,000 articles, ensuring a balanced representation across the publications. The dataset spans from 1996 to 2023.

4 Experiment 1

4.1 Subsampling

Due to the computational limitations of the analysis environment, we needed to subsample the original dataset. Our goal was to maintain a representative subset of articles while reducing the overall data size. We devised a subsampling strategy to ensure

a balanced representation across newspapers and years.

For each newspaper and year combination (excluding the years 1996 and 1997 due to data availability limitations), we initially aimed to include 200 articles. However, there were some instances where the original dataset had fewer than 200 articles for specific combinations, such as the absence of articles for the year 2006 in the French newspaper HEU. Therefore for each combination we had a minimum of 0 and a maximum of 200 articles. By employing this approach, we maintained a fair distribution of articles across different time periods and newspapers while addressing data imperfections.

After subsampling, we obtained a subset of 26,000 German articles and 19,400 French articles. These subsamples served as the basis for further data preprocessing, topic modeling, and subsequent analyses.

4.2 Data Preprocessing

To reduce the computational cost and improve the efficiency of the analysis, we applied Stopwords removal. Stopwords, common words that do not carry substantial semantic meaning, were removed using pre-trained models from Spacy, specifically de-core-news-md for German and fr-core-news-md for French.

4.3 Topic Modeling

We employed BertTopic, a transformer-based model from Huggingface, for topic modeling. The model utilizes BERT for text embedding, UMAP for dimension reduction, and HDBSCAN for clustering. By combining these techniques, we aimed to extract meaningful topics and identify clusters of related articles in the dataset.

4.4 German Topic for Bank

The German topic for "Bank" consisted of the following keywords: "bank - kunde - credit - suisse - finma - mrd - aktie - geld - grossbank - kantonalbank - institut - franken - banking - prozent - schweizer." These keywords reflect the key aspects and entities related to banking in the German-speaking regions of Switzerland.

4.5 French Topic for Bank

In contrast, the French topic for "Bank" consisted of the following keywords: "milliard - fonds - banque - franc - gestion - bnfice - march - million - morgan - vente - dollar - client - rsultat - actif - action." These keywords highlight the different aspects and entities associated with banking in the French-speaking regions of Switzerland.

4.6 Filtering

To focus on the most relevant articles, we applied two filtering steps. Firstly, for each newspaper, we selected the most frequent topics, including the "Bank" topic, as they are likely to capture the dominant subjects. Interestingly, while the "Bank" topic was among the most frequent topics for German newspapers, it did not rank as one of the top topics for the French newspapers. Secondly, we retained only those articles with a probability of belonging to their assigned topic higher than 50 percent.

After data preprocessing and filtering, we obtained a final dataset comprising 3,107 French articles and 3,015 German articles. These articles represent the most salient and relevant contributions to the analysis of banking coverage in Swiss local newspapers.

5 Results

In experiment 1, the effectiveness of BERTopic in identifying news articles related to banking topics was demonstrated. Figure 1 and Figure 2 provide clear visual evidence of the varying number of articles over time, enabling the identification of years when banking-related topics received greater prominence in Swiss newspapers. Additionally, it is noteworthy that German newspapers exhibit a higher frequency of articles on banking compared to their French counterparts.

Interestingly, despite the availability of numerous topics, banking as a subject is not covered as frequently as other topics in both languages. This finding suggests that banking issues may not enjoy the same level of popularity in Switzerland, which is unexpected.

6 Experiment 2

In order to mitigate the potential loss of articles pertinent to our study's focus on banking, despite not having been initially considered due to the constraints of our subsampling strategy, we adopted an alternative filtering approach utilizing Term Frequency-Inverse Document Frequency (TF-IDF). This method afforded us the opportunity to prioritize and retain articles most relevant to our

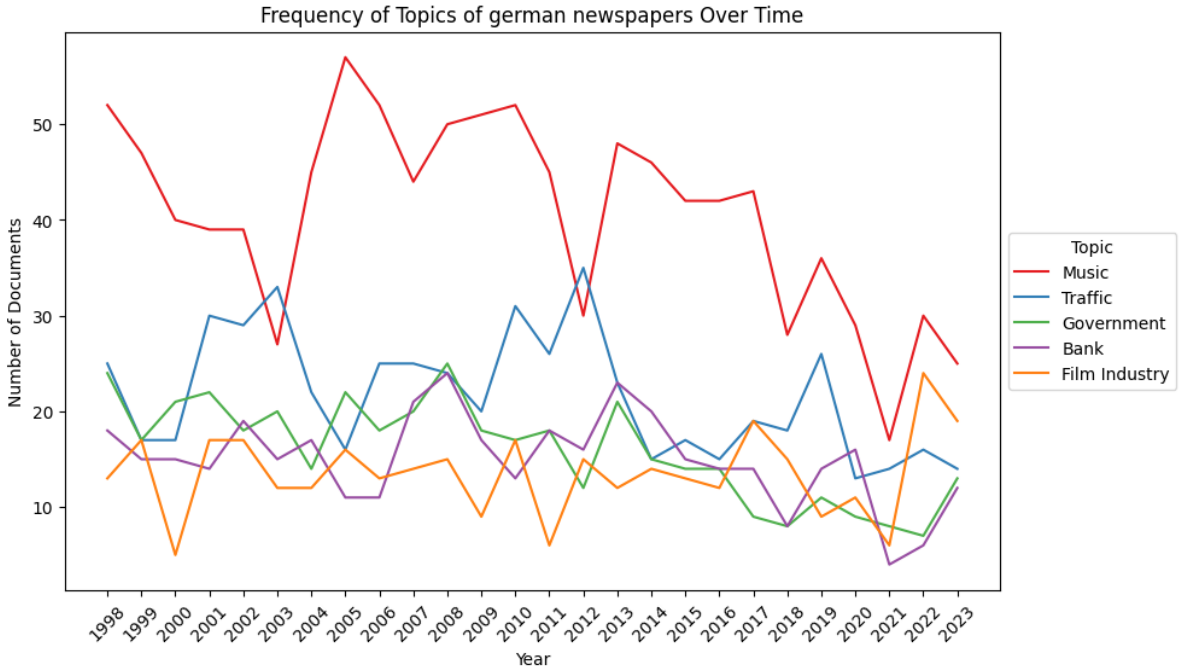

Figure 1: Frequency of different Topics over the years in German newspapers

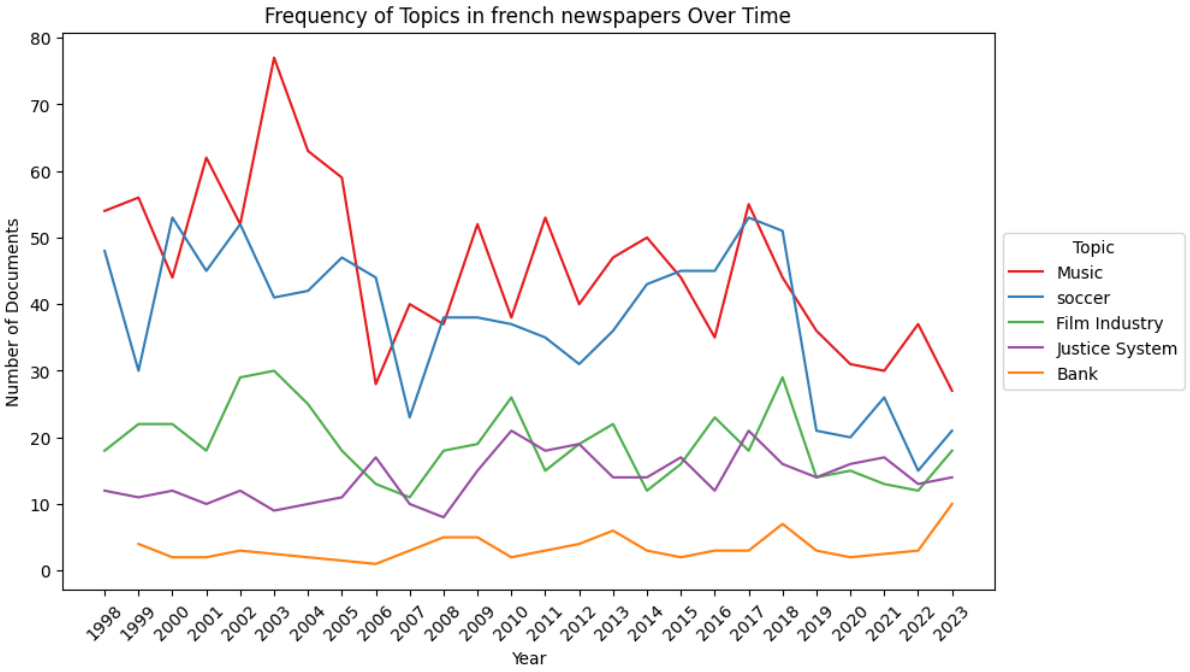

Figure 2: Frequency of different Topics over the years in French newspapers

topic of interest, ensuring a robust representation of banking-related content in our subset of data.

6.1 TF-IDF Subsampling

The refinement of a large dataset of 900,000 articles to focus on banking-related content involved multiple stages, applying TF-IDF filtering for significant refinement. The first step was text preprocessing, which converted all text into strings and removed special characters, leaving only alphanumeric characters and spaces. To ease keyword search, all text was then converted to lowercase.

The dataset was subsequently categorized based on the language of origin into German

and French articles, resulting in two separate dataframes containing 500,000 German articles and 400,000 French articles respectively.

The next stage involved refining the dataframes using keyword filtering. Separate lists of banking-related keywords were created for both German and French, which were then used to filter the respective dataframes. This process led to a considerable reduction in the data size, with the German dataframe reduced to 96,269 articles and the French one to 139,907 articles, ensuring that the remaining articles were more likely to contain banking-related content.

The final stage was the application of Term Frequency-Inverse Document Frequency (TF-IDF) filtering. This text mining technique helped in identifying the most relevant banking keywords. The banking keywords served as the vocabulary for the TF-IDF vectorizer that was applied to the text.

Each document's TF-IDF scores were then calculated, and a threshold was set to filter out documents with low scores. The challenge here was to strike a balance between the quality and quantity of articles remaining after filtering. Upon applying this TF-IDF threshold, the German dataframe held 41,110 articles, and the French dataframe retained 29,130 articles.

This meticulous methodology enabled fine-tuning of the TF-IDF threshold, striking a balance between the volume of data and its relevance to the banking industry. By testing various thresholds, an optimal balance was achieved, resulting in a banking-rich dataset beneficial for this study.

6.2 Data Preprocessing and Topic Modeling

In order to optimize the subsequent text analysis, we undertook rigorous data preprocessing. Notably, this involved removing stopwords, which are common words that lack significant meaning and can detract from the efficiency of analysis. For this purpose, we utilized the Spacy models, de-core-news-md for German and fr-core-news-md for French.

However, considering BERTopic's potential to gain understanding from context, we experimented with an approach where stopwords were retained. Nevertheless, we noticed that this led to the generation of topics populated heavily with stopwords, thereby diluting the value and interpretability of these topics. Hence, we reverted to stopwords removal, which proved to be more effective for our particular goal and dataset.

6.3 German Topic for Bank

Our methodology of data preprocessing and topic modeling led to the identification of two distinct, banking-related topics within the German newspaper dataset.

Topic 0 is largely centered around banking institutions and their operations. This is suggested by the presence of several key terms that are often associated with banks. For a comprehensive view of the terms that characterize Topic 0, in contrast, exhibits a strong emphasis on financial metrics and economic indicators. This is signified by a separate set of terms, which highlight the state of financial affairs and economic trends.

These topics together provide a well-rounded view of banking as portrayed in the German newspaper corpus, presenting insights from both the institutional framework and the economic landscape.

6.4 French Topic for Bank

Applying our topic modeling strategy to the French newspaper dataset, we identified one prominent banking-related topic. This topic is noteworthy for its blend of financial and institutional narratives in the banking domain.

Key terms used to characterize this topic encompass a range of financial metrics and institutional terminology, which together offer a comprehensive view of the banking sector.

This topic effectively captures the banking narrative in the French newspaper corpus, offering a comprehensive view that merges economic indicators with the functional and operational aspects of banking institutions.

6.5 Filtering

In our approach to refining the dataset, we adopted a probability-based filtering strategy. Only those articles that had a probability of belonging to their as-

signed topic greater than 50 percent were retained. This threshold was chosen to ensure a reasonable level of confidence in the topical assignment of each article, enhancing the reliability of our subsequent analyses. Surprisingly, this stringent criterion did not result in the removal of any articles from either the German or French datasets. All articles in our subsamples exceeded the 50

This unexpected result indicates a high level of topic coherence within the generated topics. In other words, the topics were well-defined and distinct, such that each article was clearly associated with its assigned topic. This high degree of topic coherence is a positive indicator of the quality and interpretability of the generated topics, lending further confidence to our topic modeling strategy.

7 Results

Experiment 2 yielded more comprehensive findings regarding the topics covered in Swiss newspapers. Through topic modeling, a total of 1,479 articles on banking were identified in German newspapers, whereas approximately 681 were found in French newspapers. This disparity suggests that French newspapers cover banking topics to a lesser extent compared to German media.

Figure 3 presents a timeline displaying the cumulative number of articles per year for both languages. The similarity in the slopes of both curves indicates the prominence of banking issues within specific years, ruling out randomness. A notable increase in the number of articles occurred after 2001, coinciding with the stock market downturn in 2002. Following a brief recovery period, the financial crisis in 2008 resulted in the highest number of banking-related articles in Swiss newspapers during the early years of the 21st century. This period witnessed various scandals and issues involving Swiss banks, garnering international attention. Subsequently, after 2015, the coverage of banks in news diminished. The year 2022 witnessed a notable resurgence in the number of articles in German newspapers, with the article count almost reaching the levels observed in 2011. This increase in coverage could potentially be attributed to the downfall of Credit Suisse, a prominent Swiss bank.

Figure 8 illustrates that NZZ is the most prolific in publishing articles on banking among all German newspapers, with over twice the number of articles compared to BZ, which had the fewest articles within the identified topic. However, it is worth noting that all media outlets addressed banking topics to some extent. Among French newspapers, TPS stands out as the most prominent, publishing nearly six times more articles on banking than the other identified newspapers as seen in Figure 9. The remaining newspapers have limited coverage of banking topics.

8 Conclusion

Our findings underscore that the coverage of banking issues in local newspapers has evolved over time, with significant events like the financial crisis having a clear impact on the volume of relevant articles. Interestingly, both German and French-language newspapers reflect similar trends over time, indicating that major events in the banking sector affect all regions of Switzerland. However, German-language newspapers provide more extensive coverage of banking topics than their French-language counterparts.

Considering our limitations in processing the data on our dask cluster, we found that the use of TF-IDF in experiment 2, yielded better results. By focusing primarily on articles directly related to banking, we achieved more satisfying outcomes in the Topic Modeling using BERT. The topics identified by BERT in experiment 2 were closely aligned with banking, and the interrelatedness of the topics was stronger compared to the topics in experiment 1.

8.1 Outlook

To further enhance our experiments, we could allocate more resources to the subsampling step and explore alternative techniques such as Keyword extraction. Extracting keywords from the bank-related articles may enable us to improve the results of TF-IDF or even replace it entirely.

Additionally, we could delve deeper into the bank-related articles identified in experiment 2. By employing keyword extraction or conducting additional topic modeling, we can identify specific events associated with banking in certain years. This further analysis would provide valuable insights into the temporal dynamics and specific occurrences within the realm of banking.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

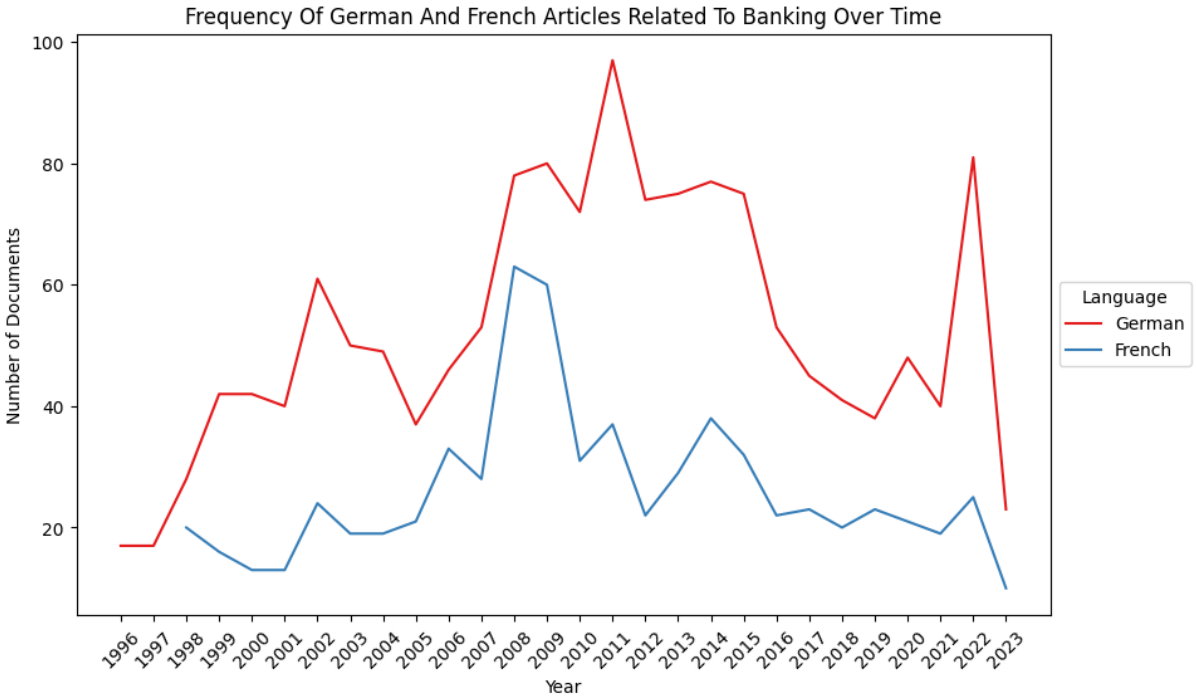

Figure 3: Frequency of bank-related articles in German newspapers over the years

Farshid Balaneji and Dietmar Maringer. 2022. Applying sentiment analysis, topic modeling, and xgboost to classify implied volatility. In *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 1–8. IEEE.

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

Weisi Chen, Fethi Rabhi, Wenqi Liao, and Islam Al-Qudah. 2023. Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study. *Electronics*, 12(12):2605.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Bayode Ogunleye, Tonderai Maswera, Laurence Hirsch, Jotham Gaudoin, and Teresa Brunsdon. 2023. Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2):797.

A Appendix

A.1 Experiment 1

Additional Charts related to 4:

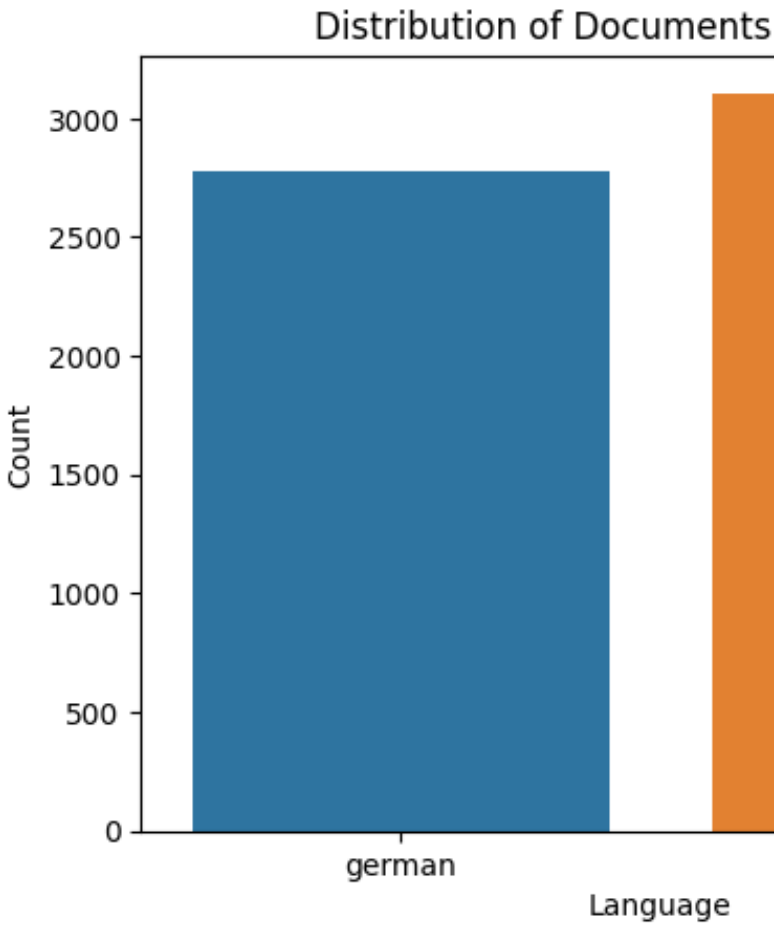

Figure 4

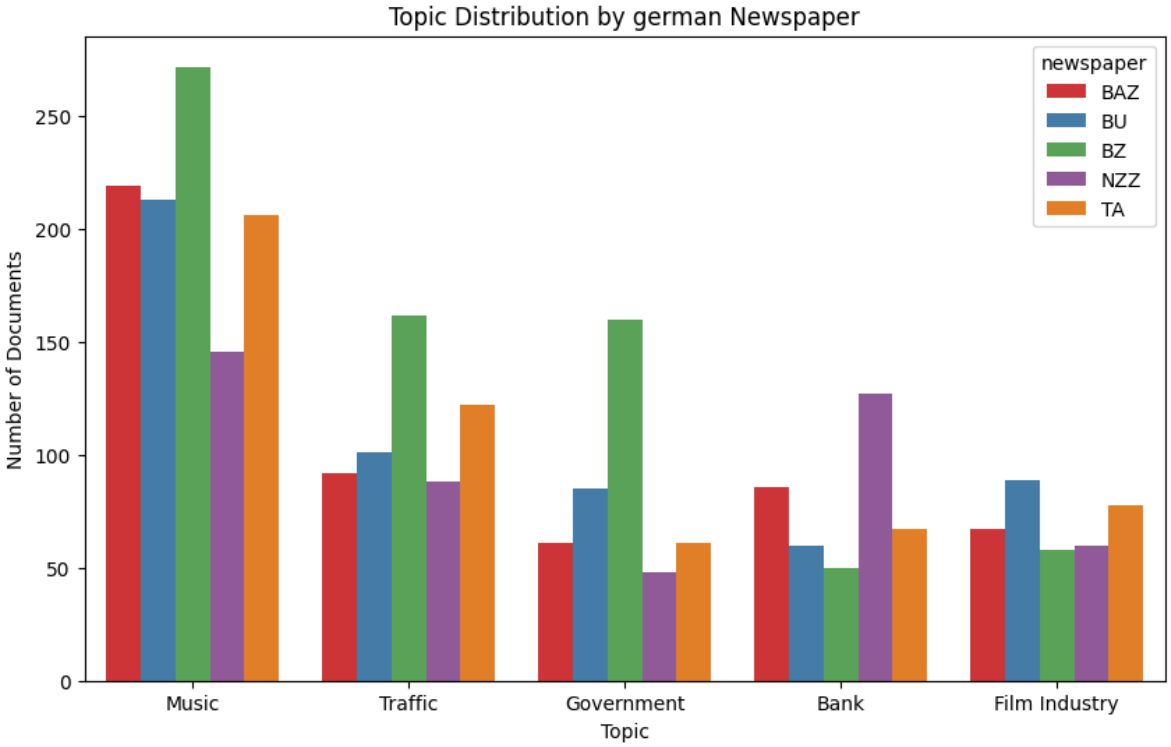

Figure 5

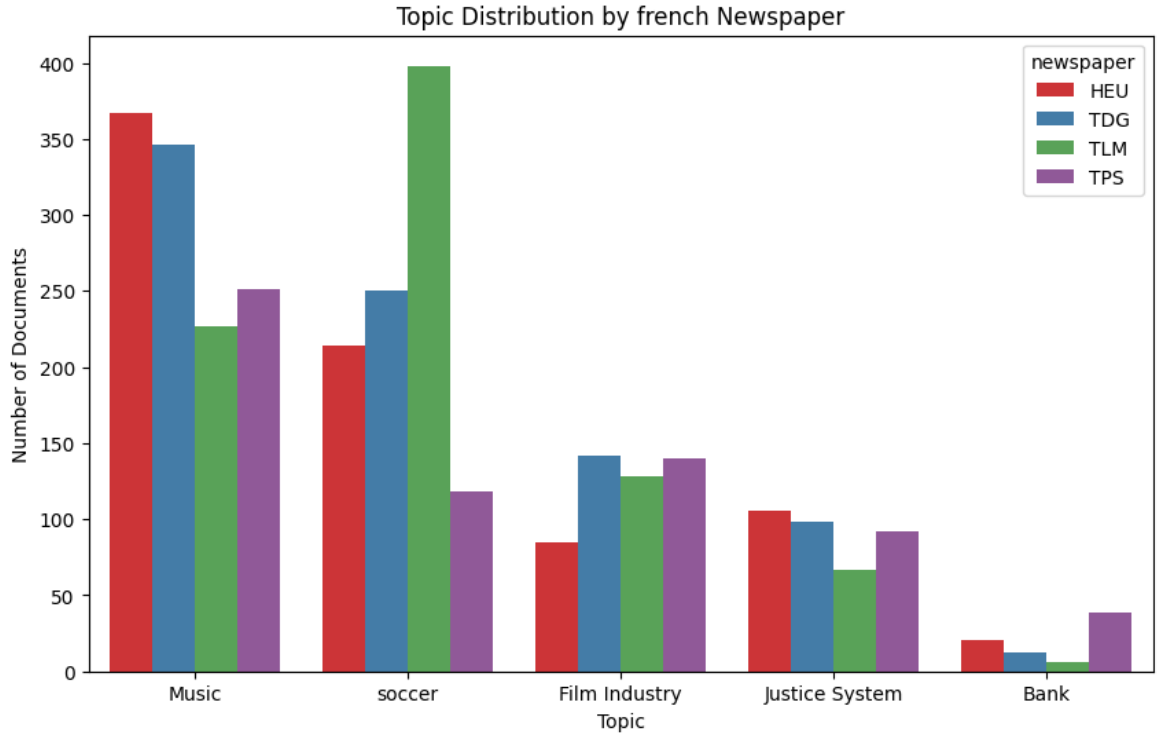

Figure 6

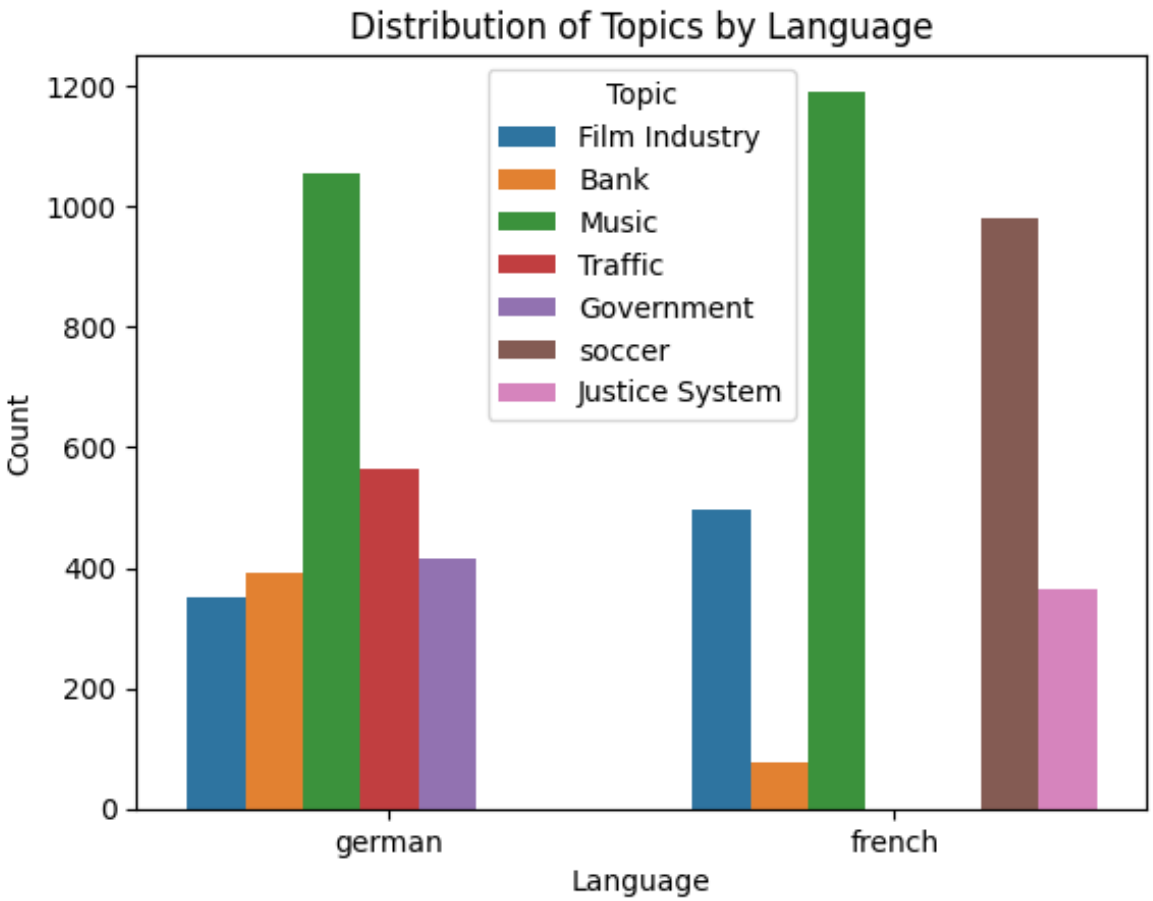
Distribution of Topics by Language
Topic
Film Industry
Bank
Music
Traffic
Government
soccer
Justice System
Count
1200
1000
800
600
400
200
0
german
french
Language

Figure 7

A.2 Experiment 2

Additional Charts related to 6:

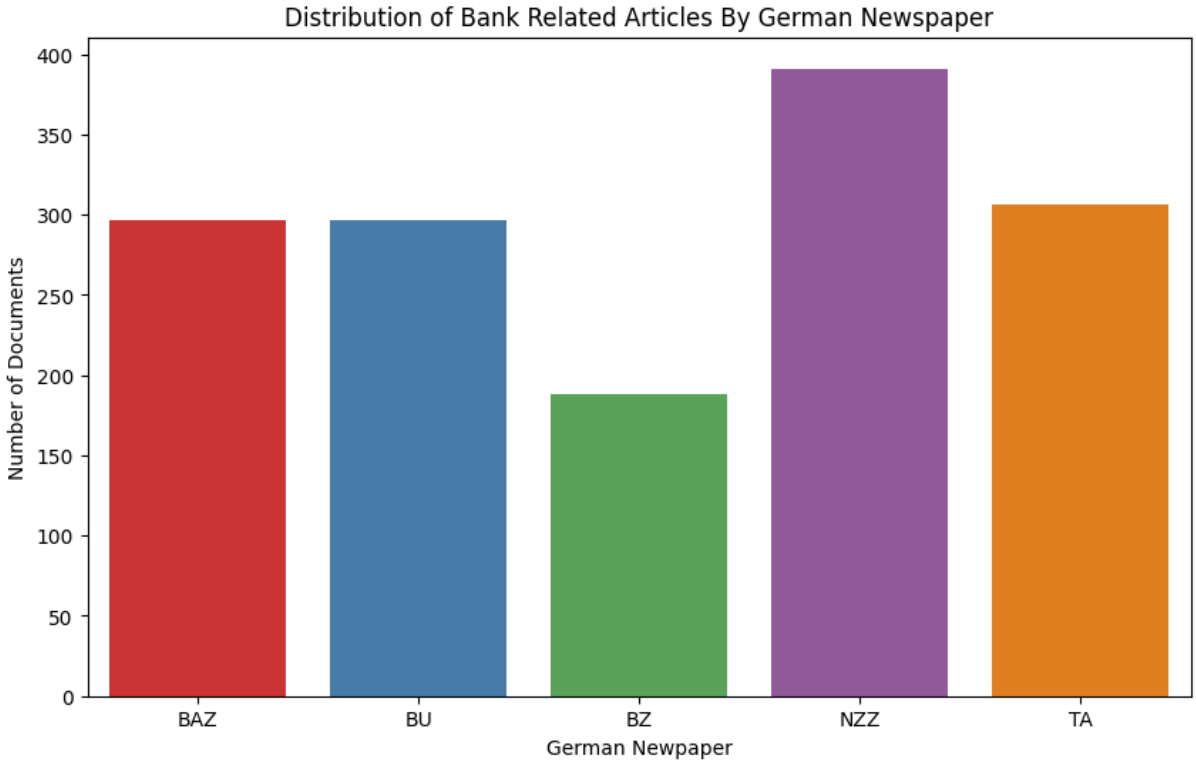

Figure 8

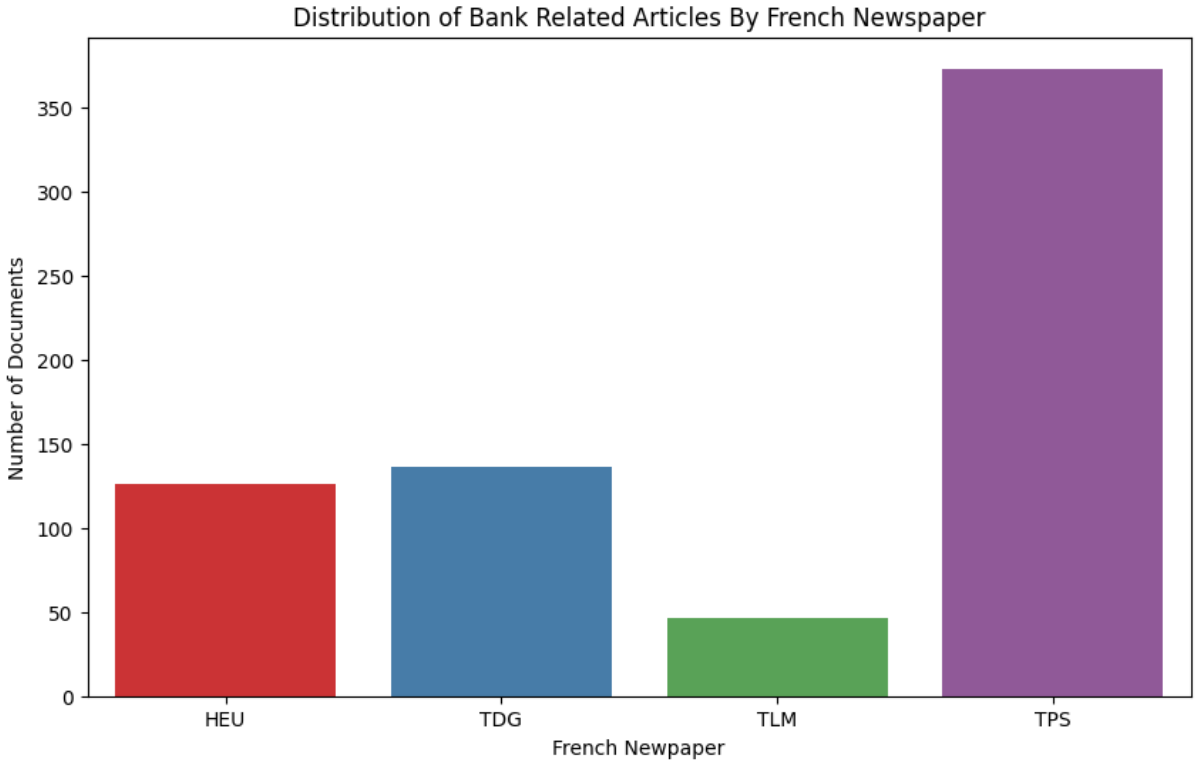

Figure 9

A.2.1 Banking Keywords French

banques, raiffeisen, bancaire, banque, ubs, postfinance, cs, zkb, banquedinvestissement, banqueprivee, secretbancaire, blanchimentdargent, baleiii, banquecentrale, bce, snb

A.2.2 Banking Keywords German

raiffeisen, bankenkrise, kantonalbank, fintech, ubs, credit, finanztechnologie, postfinance, bankenregulierung, cs, bankkrise, finanzkrise, banken, bankwesen, finanzinstitut, bank, zkb, kredit, investmentbank, privatbank, finanzprodukte, bankgeheimnis, baseliii, bundesbank, ezb, snb

A.2.3 German Topic 0 Key Terms

bank, kantonalbank, privatbank, grossbank, zürcher, zürich, nationalbank, reuters, bilanzsumme, kunden

A.2.4 German Topic 1 Key Terms

milliarden, dollar, franken, millionen, prozent, gewinn, nationalbank, bund, schweizer, bank, adani, fonds

A.2.5 French Topic for Bank Key Terms

milliards, francs, bénéfice, milliard, millions, dollars, trimestre, banque, gestion, fonds, groupe, fortune, perte, résultats, année