Project report group 3:

What Moves r/Austria?

Exploring how trends, topics, and political discussions shape engagement in Austria's largest Reddit community

Authors:	Rizvan Mukit, Bettina Münzer, Samiul Islam Sami, Jannatun Naim Sohani
Course:	Foundations of Computational Social Systems (WS 2024)
Instructor:	Karimi, Fariba, Univ.-Prof. M.Sc. Ph.D.
Institution:	Graz University of Technology, Institute of Human-Centred Computing
Date:	09-FEB-2025

Abstract

Social media platforms have increasingly become spaces for political discourse, shaping public opinion and fostering digital interactions. Among these, Reddit provides a structured environment for community-driven discussions, particularly in subreddit communities. This study examines political discussions within r/Austria, one of Austria's largest online communities, to explore user engagement patterns, trending topics, and sentiment distribution.

Using data collected from Reddit's API over a three-week period, this study employs computational techniques such as topic modeling and sentiment analysis to uncover key themes and discourse trends. The findings reveal that political discussions receive high engagement, with some categories generating significantly more interaction than others. Furthermore, sentiment analysis suggests a predominantly negative tone in political discussions, raising questions about polarization and public dissatisfaction.

Despite the insights gained, this study acknowledges limitations, including dataset restrictions, potential biases in engagement metrics, and challenges in sentiment interpretation. Future research could extend the dataset, compare multiple online platforms, and refine analytical techniques for improved accuracy. These findings contribute to the broader understanding of online political engagement and highlight the role of digital platforms in shaping political discourse.

Motivation

Social media platforms have become integral to modern political discourse, shaping public opinion and providing a space for discussion, activism, and information exchange. Among these platforms, Reddit stands out due to its unique structure, where communities (subreddits) form around specific interests, enabling dynamic user interactions. In Austria, the subreddit r/Austria has emerged as a key digital forum, accumulating over 591,000 members [1]. With Reddit's projected growth in Austria expected to reach 2.02 million users by 2028 [2], understanding how discussions unfold within this space is crucial for gaining insights into contemporary political engagement and social trends.

This study aims to explore the key topics and engagement dynamics that shape discussions in r/Austria, with a particular focus on political discourse. Political conversations on social media often serve as indicators of public sentiment, opinion polarization, and the broader socio-political climate. By analyzing trends, topics, and sentiment in r/Austria, this project seeks to answer the following questions:

- What topics dominate discussions within r/Austria, and how are they distributed across different post categories (flairs)?
- How does user engagement (measured through upvotes, comments, and posting frequency) vary across different discussion themes?
- What are the prevailing sentiments expressed in political discussions, and do they indicate a tendency toward negativity or polarization?

The relevance of this research extends beyond understanding online behavior. As digital platforms increasingly influence political narratives, analyzing Reddit data can provide valuable insights into public sentiment and political engagement patterns. This study can contribute to computational social science by demonstrating how automated techniques such as topic modeling and sentiment analysis can be applied to large-scale online discussions. Furthermore, it offers a case study of political discourse in Austria's largest Reddit community, shedding light on how citizens engage with political issues in an era of growing digital communication.

Data retrieval

Efficient and structured data collection is a crucial first step in any computational analysis of social media discourse. To ensure a comprehensive dataset for this study, a methodical approach was taken to retrieve posts and comments from r/Austria using the Reddit API. The primary objective was to extract relevant data while maintaining the integrity of discussion threads and user interactions. The following sections describe the methodology used for data collection and storage, ensuring a reproducible and transparent research process.

Data collection methodology

To gather data for this study, the Python Reddit API Wrapper (PRAW) was used. PRAW provides a structured interface to interact with Reddit's API, allowing for the extraction of posts and their associated comments.

To collect data using Reddit's API, developers are required to register as a Reddit API user. This involves creating an application under their Reddit account and obtaining authentication credentials. The generated client_id, client_secret, username, and password securely were stored in an environment file (.env). These credentials were then used to authenticate requests when interfacing with Reddit's API via PRAW.

Due to the API restrictions imposed by Reddit, only a certain number of the latest posts could be retrieved. The subreddit.new() function was used to get the maximum number of most recent posts available. The data collection is therefore limited to a period of three weeks, from December 27, 2024 to January 16, 2025. The contributions were saved starting from the current date for each day backwards.

Data structure and storage

Each day's data was stored in a separate JSON file to facilitate structured data management and later analysis. While the retrieved dataset preserved the hierarchical structure of posts and comments, it was stored in a flattened format, meaning that parent-child relationships between comments were not retained. Each post entry contained essential metadata, including a unique identifier, title, text content, author, creation timestamp, upvotes, number of comments, and assigned flair. Similarly, comments were stored with their unique identifier, text content, author information, creation timestamp, and upvote count. This structured approach ensured consistency across the dataset while enabling efficient data processing and analysis.

Data processing

Python was chosen as the programming language for this analysis as it provides extensive libraries and tools for data processing, statistical analysis and natural language processing. The programming language was also used in the course, making it an appropriate choice for consistency with the methods presented in the lectures.

The analysis of the dataset followed a structured approach, beginning with a general statistical overview before proceeding with more advanced text processing techniques. Each key step in the data processing workflow is outlined in the following sections.

Statistical analysis: engagement and flairs

A fundamental aspect of the analysis involved obtaining a statistical overview of user engagement and activity within the r/Austria subreddit. This included calculating the total number of posts, comments, and upvotes, as well as examining daily activity trends to understand fluctuations in engagement over time.

Daily statistics were computed to assess the average number of posts and comments per day, along with their standard deviations. Additionally, peak and low activity days were identified to highlight engagement patterns. The correlation between posts and comments was analyzed using Pearson's correlation coefficient, providing insights into how post frequency influences discussion activity.

Furthermore, engagement patterns across different flairs were examined. The dataset was filtered by post categories, revealing that while some flairs, such as "Memes & Humor," received the highest number of upvotes, political discussions under the "Politik | Politics" flair generated the highest volume of comments. These statistical findings informed the decision to focus subsequent analyses on the political discourse within the subreddit.

Data cleaning

To ensure the consistency and reliability of the dataset, a structured data cleaning process was applied. This involved standardizing textual data, filtering out irrelevant content, and preparing the dataset for subsequent analyses such as word cloud generation, topic modeling, and sentiment analysis.

The data cleaning process consisted of the following key steps:

- **Text normalization:** All text data was converted to lowercase to ensure uniformity and facilitate text comparison.
- **Removal of unnecessary elements:** URLs, special characters, and emojis were eliminated to reduce noise in the dataset and to eliminate extraneous content that could interfere with linguistic analysis. .
- **Tokenization and stopword removal:** The text was split into individual words, and common stopwords in both German and English were removed. A predefined list of stopwords from spaCy's language models was used, supplemented by manually identified stopwords based on frequent word analysis.
- **Handling empty entries:** Posts and comments that contained no meaningful content after preprocessing were excluded from further analysis.
- **Subset generation:** Isolating posts based on specific flairs for targeted analysis.

These preprocessing steps were crucial for improving the quality of the word cloud, topic modeling, and sentiment analysis results.

Word cloud and frequency analysis

To better understand the most frequently discussed terms within the subreddit subset with flair politics, a word cloud and word frequency analysis were conducted. These methods provide a visual and quantitative representation of the dominant themes and keywords in the dataset.

The word cloud was generated using the WordCloud library in Python, which aggregates the most commonly occurring words from the dataset and visually represents them, with larger words indicating higher frequency. A custom stopword list, derived from spaCy's German and English stopwords and supplemented with additional manually identified terms, was applied to filter out commonly used but uninformative words.

In addition to the word cloud, a word frequency analysis was performed to identify the most frequently occurring words across all posts and comments. The frequency distribution was saved as a CSV file and visualized using a horizontal bar chart, highlighting the top 20 most frequently used words.[3] This analysis helps in identifying key discussion points and recurring themes in political discourse within r/Austria.

Furthermore, insights gained from the word frequency analysis allowed for the refinement of the stopword list. Frequently occurring yet contextually insignificant words were identified and added to the custom stopword list, ensuring that future analyses would focus on more meaningful terms. This iterative process improved the quality and relevance of topic modeling and sentiment analysis.

Topic modeling

To identify the main themes within the political discussions on r/Austria, topic modeling was performed using Latent Dirichlet Allocation (LDA). This probabilistic model groups words into topics based on their co-occurrence patterns within the dataset, allowing for a structured analysis of recurring themes.

The LDA model was implemented using scikit-learn's LatentDirichletAllocation in combination with the CountVectorizer for feature extraction. The number of topics was set to four, based on experimental tuning and coherence score evaluation. Each topic was represented by the ten most relevant words, providing insight into the dominant themes within the dataset.

To ensure the validity of the topics, a coherence score was calculated using Gensim's CoherenceModel. This metric assesses the semantic similarity of words within each topic, ensuring that the identified themes are meaningful and interpretable.

Sentiment analysis

To assess the tone of political discussions within r/Austria, sentiment analysis was conducted using a BERT-based multilingual model (nlptown/bert-base-multilingual-uncased-sentiment). The selection of this model was based on the fact that discussions within the subreddit take place in both German and English. The model classifies text into five sentiment categories based on a star rating system (1 star to 5 stars). To enhance interpretability, these labels were manually mapped to sentiment categories from very negative to very positive.

The analysis process involved:

- **Preprocessing text:** The cleaned dataset was used, incorporating both post titles, selftext, and comment texts.
- **Applying sentiment classification:** The BERT model was used to classify each text segment into one of the five sentiment categories.
- **Handling text length constraints:** Since the BERT model has a maximum input length of 512 tokens, longer texts were truncated to ensure compatibility.

Results

The following section provides an overview of the main findings. Results from the word frequency analysis were incorporated to improve further text analysis and are therefore not explicitly reported here.

Data overview and flair analysis

A total of 977 posts and 47,775 comments were collected over a period of three weeks. The number of posts per day varied significantly [4]:

- Average Posts/Day 46.52 (STD: 17.25, MIN: 18, MAX: 89)
- Average Comments/Day 2 275 (STD: 926.47, MIN: 619, MAX: 4 120)

A strong correlation between the number of posts and the number of comments was observed, which was confirmed using pearson correlation: R = 0.88, p-value = 2.04E-07. [5] However, no significant correlation was found between the number of upvotes and the number of comments: R = 0.36, p-value = 7.39E-32. [6]

With 8241 unique users contributing during the analyzed period, the subreddit exhibited a high level of activity.

On Reddit, flairs can be optionally assigned to categorize posts. In the r/Austria subreddit, 100% of posts within the observed period were tagged with a flair, making it easy to analyze user engagement across different categories.

Upvotes and comments serve as further key indicators of engagement on Reddit. Upvotes represent community appreciation and influence a post's visibility, while comments reflect interaction and discussion intensity.

Four flair categories with the highest engagement were identified [7–11]:

- **Frage | Question:** This flair accounts for 40.0% of posts, 21.3% of comments, and 6.2% of upvotes.
- **Politik | Politics:** With 18.7% of posts, 36.0% of comments, and 26.1% of upvotes, it emerges as a dominant topic both in discussions and upvoting behavior.
- **Memes & Humor:** While making up 14.8% of posts and 12.3% of comments, this flair receives 60.8% of upvotes, highlighting its strong appeal to users.
- **Nachrichten | News:** Representing 7.5% of posts, 10.6% of comments, and 7.0% of upvotes, this flair holds a moderate presence within the overall discussions.

Topic modeling

With the help of the LDA, four key topics in the political discussion could be identified [12]:

1. **Austria & Economy:** Discussions about money, living costs, and geopolitical issues.
2. **Parties & Politicians:** Focus on SPÖ, ÖVP, FPÖ, Neos, Babler, Kickl.
3. **Social Concerns:** Taxes, financial burdens, societal challenges.
4. **Government & Power:** Strategies, parties, and power dynamics.

The Coherence Score (0 to 1) measures how logically consistent the identified topics are. A higher score is indicating greater consistency between words within a topic which means that the words are more strongly related to each other, making the topics more coherent and relevant. Although achieving a coherence score higher than 0.47 was not possible, this result remains meaningful given the limited dataset.

Sentiment analysis

The sentiment analysis successfully processed a total of 15657 texts from political posts and comments. The results indicate a predominance of negative sentiment, with the majority of texts rated as "Very Negative" (6362). In contrast, positive sentiment is less common, with only 2642 texts receiving a "Very Positive" rating. The distribution of sentiments is as follows: Very Negative: 40.6%, Neutral: 24.2%, Very Positive: 16.9%, Positive: 6.3%, and Negative: 12% [13]. This suggests that while some posts and comments express positivity, the overall discourse in the political context tends toward critical sentiment.

Conclusion

This study highlights the significance of r/Austria as a digital space where political discourse and engagement take place. Through the analysis of trends, topic modeling, and sentiment evaluation, key insights into the dynamics of online political discussions were uncovered. The findings suggest that political content generates substantial interaction, with engagement varying across different post categories. The dominance of negative sentiment in political discussions also points to potential concerns regarding polarization and public dissatisfaction.

To evaluate the reliability of these findings, it is important to consider the limitations of the dataset and methodology. The study relied on a three-week sample of posts and comments, which may not fully represent long-term trends in r/Austria. Additionally, user participation on Reddit is self-selecting, meaning that certain demographics or perspectives may be over- or underrepresented. The sentiment analysis, while leveraging a BERT-based model, is inherently constrained by the nuances of language and context, particularly in multilingual discussions. Topic modeling, though effective in identifying thematic structures, depends on pre-set parameters that influence the coherence and interpretability of topics.

Despite these limitations, the methodological approach provides a meaningful glimpse into the nature of political engagement within the subreddit. Future research could enhance reliability by extending the data collection period, incorporating additional linguistic preprocessing techniques, and comparing sentiment and engagement trends across multiple online platforms.

Critique

While this study provides valuable insights into the nature of political discussions in r/Austria, several limitations must be acknowledged. First, the dataset is inherently limited in its timeframe, covering only three weeks of activity. Political discussions often fluctuate based on external events, such as elections or major policy changes, meaning that a longer-term dataset could yield different insights.

Second, the analysis assumes that engagement metrics like upvotes and comments directly reflect user interest and sentiment, but these metrics can be influenced by Reddit's internal algorithms and voting biases. Highly engaged posts may receive more visibility, creating a feedback loop that amplifies certain discussions while muting others. This effect could skew the observed trends and misrepresent the true diversity of opinions.

Third, while the sentiment analysis model used in this study is state-of-the-art, sentiment classification remains a challenging task, especially for multilingual and informal discussions. Sarcasm, irony, and nuanced language may not be accurately captured, leading to potential misclassifications. Additionally, topic modeling using LDA has inherent limitations, as the number of topics and their coherence scores depend on parameter tuning, making the results somewhat subjective.

Alternative explanations should also be considered. The prevalence of negative sentiment in political discussions may not necessarily indicate polarization but could reflect a general dissatisfaction with political institutions, a phenomenon observed in many online forums. Furthermore, high engagement in political posts could be attributed to the controversial nature of political discussions rather than genuine civic interest.

Addressing these limitations in future research would enhance the robustness of findings. Expanding the dataset, employing more sophisticated machine learning models, and incorporating qualitative methods such as manual content analysis could provide a more comprehensive understanding of political discourse on Reddit.

References and links

[1] Reddit Netherlands B.V. reddit.com Website. URL: https://www.reddit.com/r/Austria/ (visited on 2025-01-15)

[2] Statista GmbH. "Forecast of the number of Reddit users in Austria from 2020 to 2028 (in millions)." - statista.com Website. URL: https://www.statista.com/ (visited on 2015-01-08)

[3] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/word_frequencies_bar_chart.png

[4] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/daily_posts_comments_overlay.png

[5] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/daily_posts_vs_comments.png

[6] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/post_upvotes_vs_comments.png

[7] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/all_flairs_number_posts.png

[8] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/all_flairs_number_comments.png

[9] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/specific_flairs_distribution_posts.png

[10] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/specific_flair_distribution_comments.png

[11] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/specific_flairs_only_posts_pie_upvotes.png

[12] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/ topics.json

[13] https://github.com/BettyInnovates/redditAustriaAnalysis/blob/main/report/results/sentiment_distribution_highlight_max.png