From Posts to Patterns

Tracking Swiss Economic Sentiment on *Reddit*'s r/Switzerland

Submission Date: August 14, 2025

Stefan Freitag

Mail: stefan.freitag@unibas.ch

Mat No.: 16-059-529

Digital Humanities & Media Studies

Supervisors:

Prof. Dr. Lukas Rosenthaler (lukas.rosenthaler@unibas.ch)

Prof. Dr. Peter Fornaro (peter.fornaro@unibas.ch)

University of Basel

Digital Humanities Lab

Switzerland

University of Basel

Digital Humanities Lab

Table of Contents

List of Figures

Statement on the Use of AI Tools

During the course of writing this thesis and the development of its associated code, AI-based tools (specifically ChatGPT, DeepSeek and Overleaf's Writefull) were used to assist with paraphrasing, language correction/suggestion, and coding-related suggestions. These tools supported the textual clarity and code development. All suggestions generated by AI were critically reviewed, evaluated, and, where appropriate, modified or entirely discarded by the author to ensure academic accuracy and integrity.

1 Introduction

In recent years, online conversations have become a valuable lens for understanding public opinion. Whether reacting to breaking news or market instability, people increasingly share their thoughts in real time on social media. These expressions of sentiment are not just personal, they can collectively signal broader societal moods, sometimes aligning with, or even anticipating measurable economic trends.

While researchers have explored this connection on platforms like *Twitter/X* and the Chinese platform *Weibo*, the Swiss context remains understudied.

Among the many platforms where such discussions unfold, *Reddit* offers a distinctive combination of both global reach and localized focus. Unlike *Twitter/X*'s brevity or *Weibo*'s regional confinement, *Reddit* encourages longer exchanges within topic-specific communities. These communities range from broad, global forums to highly localized spaces such as r/Switzerland, where users discuss domestic topics, including the economy. While *Reddit* has received some scholarly attention for economic sentiment analysis, most studies focus on theme-specific communities like r/wallstreetbets.

This thesis broadens the picture by investigating **whether sentiment expressed in economy-focused r/Switzerland posts reflects real-world developments in the Swiss economy, and whether it could serve as a predictive signal for future trends**. To do so, it combines multiple natural language processing (NLP) techniques: sentiment analysis using both VADER and FinBERT, named entity recognition with SpaCy, and topic modeling with BERTopic. The analysis is conducted on a year-by-year basis from 2020 to 2025, comparing sentiment patterns directly against key Swiss economic indicators such as GDP growth, inflation, stock market performance, and consumer sentiment.

By joining online discourse and official economic data, this work not only evaluates the representativeness of *Reddit* sentiment in the Swiss context but also contributes to the broader understanding of how social media can complement traditional economic measurement. The following chapters build towards this aim:

Chapter 2 reviews relevant literature involving *Reddit* as a research subject and the application of NLP-based approaches on social media to gain economic insights. **Chapter 3** builds the theoretical foundation, discussing *Reddit* as the data source, as well as data collection, storage, and analysis. **Chapter 4** details the methodological implementation of

what is outlined in Chapter 3. **Chapter 5** presents the results produced by this approach. **Chapter 6** discusses these results in relation to Swiss economic indicators, highlighting key trends and limitations. Finally, **Chapter 7** concludes the thesis by summarizing the main findings, their alignment with previous literature, and suggesting avenues for future research.

2 Literature Review and Related Work

This chapter reviews key research on *Reddit* and natural language processing (NLP) applications for economic insights, establishing the foundation and gap this thesis seeks to address.

2.1 Prior Research on *Reddit*

Reddit's unique structure and community dynamics (both of which will be explained in a later section) have made it an increasingly common subject of academic exploration, despite its lesser-known status in mainstream discourse. Due to the diversity of content available on *Reddit*, researchers have used it to examine a multitude of topics, including the differences between pro-Trump and pro-Clinton communities during the 2016 presidential elections (Jungherr et al., 2022), its role in the distribution of false information (Weld et al., 2021) or within the context of community moderation processes (Gilbert, 2020).

Regarding data collection and analysis tools, most data for research purposes is generated using the official *Reddit* application programming interface (API). These datasets are often analyzed using computational methods such as NLP or topic modeling techniques based on Latent Dirichlet Allocation (LDA). In terms of dataset size, most collections consist of between 100 and 1,000 posts (Proferes et al., 2021).

The user demographic of *Reddit* has also been the focus of prior research, with findings indicating that the platform's user base is typically young (in their 20s), predominantly male, and primarily English-speaking (Gjurković et al., 2021). This demographic skew should be kept in mind when interpreting findings derived from *Reddit*-based data.

2.2 Research on NLP for Economic Insights

A growing body of research has applied NLP techniques to extract economic insights from social media platforms across various countries. While studies have examined several platforms, there remains a lack of research focused specifically on the Swiss economy. For instance, sentiment expressed on *Weibo*, a Chinese microblogging site, has been used to forecast economic trends in China. In this context, sentiment-based models were shown to outperform traditional market indicators (Pan and Li, 2015).

Other studies have focused on global platforms such as *Twitter/X*. Given the worldwide economic impact of the COVID-19 pandemic, researchers have analyzed sentiment within *Twitter/X* data to investigate topics like its influence on Singaporean housing prices. This approach has been shown to produce more reliable predictions compared to models that do not take sentiment into account (Durai and Wang, 2023). Additionally, other researchers have used *Twitter/X* sentiment to construct economic indices, which have mirrored real-world economy-related events and exhibited correlations with market indicators (Fano and Toschi, 2022). Based on these prior studies, economic sentiment expressed on social media platforms can be expected to reflect real-world economic trends.

Recent research has also examined *Reddit* as a source of economic sentiment. Much of this work focuses on economy-specific communities centered around stock trading, such as r/wallstreetbets, with the aim of predicting certain stock prices or refining trading strategies (Machavarapu, 2022; K et al., 2024). Other studies have analyzed comments from across the entire platform to forecast broader market indicators, such as the performance of the American S&P 100 index (Chen and Ma, 2024). Overall, these studies suggest that *Reddit*-based sentiment analysis holds considerable potential and is a promising area for further exploration. Building on these findings, some researchers have developed practical tools that apply sentiment analysis to *Reddit* data. For example, *Reddiment* (Bauer et al., 2022) offers a web interface for visualizing correlations between online sentiment and stock price data, demonstrating growing interest in making these insights accessible for applied use.

2.3 Research Gap and Motivation

Although both *Reddit*-based research and the use of natural language processing to extract economic insights from social media have been explored, studies applying these methods

to the Swiss economy are scarce. One notable exception examines sentiment in Swiss news media, suggesting that during times of economic volatility, such sentiment data could potentially enhance forecasting accuracy (Bieri, 2023). However, the application of similar sentiment analysis techniques to platforms like *Reddit*, in a Swiss setting, remains largely unexplored.

To address this gap, the present thesis investigates the potential of sentiment analysis on *Reddit* data (specifically from the r/Switzerland community) to generate economic insights within the Swiss context. By applying NLP techniques to social media discourse, this research aims to complement existing economic indicators and contribute to a deeper understanding of how digital platforms reflect or even anticipate economic developments in Switzerland.

3 Theoretical Foundation

This chapter lays the theoretical foundation by introducing and defining the key concepts and tools used, following a top-down approach: beginning with broad, high-level concepts such as social media and *Reddit*, then narrowing the focus to more specific technical elements like application programming interfaces (APIs) and natural language processing techniques. This structure reflects the workflow of the research itself: starting with the selection of a data source, followed by data collection, storage, and finally, analysis. Each section is designed to provide the necessary background to understand the methodological decisions and analysis techniques used in later chapters.

3.1 Data Source - Social Media and *Reddit*

3.1.1 Social Media in general

Before going into the finer details and specifics of *Reddit* as the chosen platform, it should first be defined what exactly a **social media platform/network** is and what differentiates it from other online platforms and applications.

Defining this term necessitates explanation of two closely associated concepts: **Web 2.0** and **User Generated Content**. Web 2.0 can be seen as a foundation without which social media could never really have emerged and can be defined as follows: "Web 2.0 is a term [...] to describe a new way in which software developers and end-users started to

utilize the World Wide Web; that is, as a platform whereby content and applications are no longer created and published by individuals, but instead are continuously modified by all users in a participatory and collaborative fashion" (Kaplan and Haenlein, 2010, pp. 60–61).

Building on this, "User Generated Content (UGC) can be seen as the sum of all ways in which people make use of Social Media. The term [...] is usually applied to describe the various forms of media content that are publicly available and created by end-users" (Kaplan and Haenlein, 2010, p. 61). This definition gives a good overview, but does not provide a direct answer to a very important question: what exactly is User Generated Content?

The Organisation for Economic Co-operation and Development (OECD) defines User Generated Content as material that meets three conditions: it must be shared publicly (e.g., on social platforms), created with individual creative input, and developed outside of traditional professional content production environments (Vickery and Wunsch-Vincent, 2007).

Now that the foundational terms of Web 2.0 and User Generated Content have been explained, both definitions can be combined to come to a possible conclusion as to what social media are: "[...] Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (Kaplan and Haenlein, 2010, p. 61).

Of course, this definition given by Kaplan and Haenlein is not the only possible way of explaining what social media means, but provides a good general overview and is therefore sufficient as a foundation for this thesis going forward, especially since it is not about the history of social media or its possible impacts on society, but rather seeks to apply methods of Digital Humanities to a specific example, *Reddit*.

3.1.2 *Reddit* as a Data Source

Founded in 2005 by Steve Huffman and Alexis Ohanian, *Reddit* is a platform for user-generated content and web aggregation (Macale, 2011; Carlucci, 2024). It was selected as the data source for this thesis due to its unique subcommunity-based structure and data accessibility.

Figure 1: The icon of *Reddit*, depicting its official mascot *Snoo*. Source: Reddit Inc., 2025a

Reddit operates by organizing its user base into interest-based communities, called **subreddits**. These subreddits are named using the "r/" prefix (e.g., r/gaming, r/pics), and as of 2025, over 100,000 active subreddits cover a vast array of topics. Within these communities, users can post text, links, images, or videos, which other users can comment on and upvote or downvote, influencing visibility. With over 100 million unique daily users and more than 20 billion cumulative posts and comments, *Reddit* is one of the most active platforms online (Reddit Inc., 2024).

Rather than forming a single continuous network, *Reddit*'s users and content are distributed across many smaller, topic-specific communities. This is what differentiates *Reddit* the most from other social media platforms:

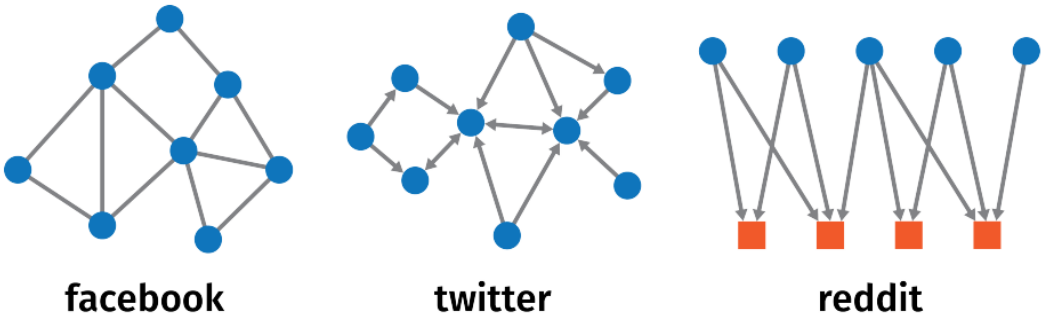

Figure 2: The structure of *Reddit* as opposed to other social networks. Users are represented by blue dots, orange squares represent communities. Source: Carlucci, 2024, p. 2

This structure makes *Reddit* a so-called **"bipartite network"**, divided between users and the communities they interact in. Furthermore, the way users of *Reddit* interact with content and each other is also different from other platforms: "[...] interactions on *Reddit* are considerably more impersonal, as almost all users operate under a pseudonym [...]" (Carlucci, 2024, p. 2).

Despite these structural differences, *Reddit* still qualifies as a social media platform under the earlier definition: it builds on Web 2.0 principles and facilitates user-generated content and interaction.

3.2 Data Collection

3.2.1 Application Programming Interfaces (APIs)

In research involving data collection from social media platforms, accessing the necessary data is often the first major challenge. To help this process, researchers commonly rely on **application programming interfaces** (APIs).

APIs ease software development by providing predefined methods for different applications to interact with each other, for example with the goal of sharing data between them. By abstracting the details of the underlying implementation, they allow developers to integrate complex functions without building them from scratch (Mudassir and Mushtaq, 2024). They can be imagined as a postman responsible for transporting data from A to B. In the case of this thesis, having an API available is very helpful as it enables data collection using its already existing tools rather than having to develop own ones.

Over the years, different types of APIs have been developed, among them are:

- APIs based on **GraphQL**, a query language that is used to address some of the limitations of older and more classical approaches (Quiña-Mera et al., 2023).
- **SOAP** (Simple Object Access Protocol) APIs, which use the **XML** (eXtensible Markup Language) format for data exchange (Soni and Ranga, 2019).

By far the most widely adopted and popular API type is **REST** (Quiña-Mera et al., 2023; Mudassir and Mushtaq, 2024). REST stands for "Representational State Transfer", which was coined by Roy Fielding based on his description of the underlying architecture of the Web (Fielding, 2000). Such **REST APIs** are built around the **HTTP** (Hypertext Transfer Protocol) and follow the principles outlined by Fielding. Web services that implement these principles are commonly described as **RESTful** (Masse, 2011). The basic operation of such APIs is illustrated in Figure 3 below:

Figure 3: Exemplary working principle of a REST API: The client communicates with the API via HTTP methods such as `GET` and `POST`. The API communicates back and forth with the server and returns data to the user as JSON (JavaScript Object Notation). Source: Own illustration

Although REST is widely adopted, it has certain limitations in terms of data retrieval. These include the complexity of executing multiple HTTP requests to gather related data, the risk of retrieving more information than necessary (**over-fetching**), and situations where insufficient data is returned, requiring additional requests (**under-fetching**) (Quiña-Mera et al., 2023).

3.2.2 *Reddit*'s API

In the case of *Reddit*, its official API follows this REST architectural style, evident from its use of standard HTTP methods such as `GET`, `PATCH`, and `PUT`. The API returns responses in JSON format, often structured using *Reddit*'s custom **Listing** format, which provides pagination metadata and items such as posts or comments (Reddit Inc., 2025b).

This API provides numerous ways (called **endpoints**) to create posts/comments or search through them. For the purposes of this study, the endpoint `GET [/r/subreddit] /search` is the ideal one to use. It allows clients to perform keyword-based searches within a specified subreddit and supports query parameters for filtering results by relevance, recency, or popularity. The endpoint returns a JSON-formatted paginated list of matching posts, enabling automated content collection from targeted subreddit communities (Reddit Inc., 2025c).

Although it is entirely possible and feasible to interact with the *Reddit* API directly through raw HTTP requests and responses, the data collection and analysis pipeline will be built using **Python**. In this context, the **PRAW** library is particularly useful. PRAW stands for "Python Reddit API Wrapper" and provides a more user-friendly, Python-based way to interface with the *Reddit* API, greatly simplifying interactions by abstracting away the low-level HTTP details. Additionally, **AsyncPRAW** offers an asynchronous version of this library, which also allows efficient handling of multiple API requests concurrently through asynchronous programming (Boe, 2023a; Payne, 2024a).

When using the API either directly or via PRAW/AsyncPRAW, adherence to its terms and conditions is mandatory. Since this thesis does not intend to use *Reddit* data to train artificial intelligence/machine learning models, explicit user permission is not needed to collect data (Reddit Inc., 2023). To maintain user privacy, all data will be anonymized, removing the username of the person that created it before saving.

Furthermore, compliance with the rate limits set by the API is crucial. Specifically,

the API allows up to 100 queries per minute per client ID. To accommodate brief spikes in traffic, this limit is calculated as an average over a rolling 10-minute period (Reddit Inc., 2025d).

Lastly and perhaps most importantly, *Reddit*'s API comes with a major inherent downside: a retrieval limit of a maximum of 1,000 items per query. This restriction stems from the Listing format used by the API to return data. It is poorly documented in the official API documentation, but has been widely acknowledged by developers and community members, and can be found in the official reference for both PRAW and AsyncPRAW (Boe, 2023b; Payne, 2024b). As a result, for any given query, only 1,000 unique items can be returned. However, this limitation can be slightly alleviated by performing multiple successive queries with a different sorting method ("new", "top", etc.), which yield different results. This can retrieve additional items, though overlaps between sorting methods are likely.

3.2.3 Alternative Methods - Web Scraping and Pushshift

Despite the limitations and constraints of the official *Reddit* API, its use remains preferable to alternative methods like **web scraping**. Scraping is an automated method of extracting data from the Web (also called "data mining"), such as parsing HTML files using tools like the **Beautiful Soup** package for Python (Lotfi et al., 2021). Unlike official APIs, which (like in *Reddit*'s case) often impose rate limits or authentication requirements, scraping can bypass these restrictions, but also comes with its own downsides: its reliance on parsing website structures makes it inherently brittle compared to APIs, as even minor layout updates can break data extraction (Alrashed et al., 2020). This fragility can render scraping-based data collection nonrepeatable and its results entirely unverifiable. For this reason, this thesis avoids web scraping entirely.

In the context of *Reddit*, **Pushshift**, a project that collects and exposes *Reddit* data via its own API, offers distinct advantages over the official API, including higher query limits and streamlined access to historical data (Baumgartner et al., 2020). However, as of writing this thesis, access to the Pushshift API has been restricted to moderators of *Reddit* and is limited to explicit moderation purposes only (NCRI / Pushshift, 2025). While archived Pushshift data dumps remain available through various third-party repositories and searchable via community-built APIs, these tools rely on static datasets that lack any

official support and vary in stability as well as recency. Furthermore, relying on unofficial services introduces major concerns around reproducibility and compliance with *Reddit*'s data usage policies. Consequently and despite its comparatively limited functionality, the official *Reddit* API was selected in order to collect the necessary data for this research.

3.3 Data Storage

After having collected the desired data, choosing appropriate data storage techniques is critical for ensuring both data integrity and accessibility. This section outlines common approaches suitable for the proposed workflow: **relational databases** and **serialization formats**.

3.3.1 Relational Databases

Relational databases, introduced by E. F. Codd in 1970, store and retrieve data using structured tables. These tables, or **relations**, organize data into columns (defining categories) and rows (holding unique entries). Constraints and predefined domains ensure data integrity. Unlike rigid hierarchies, relational databases allow flexible data access through table relationships, enabling multiple perspectives on the same dataset (Jatana et al., 2012). An illustration of their working principle can be found below in Figure 4:

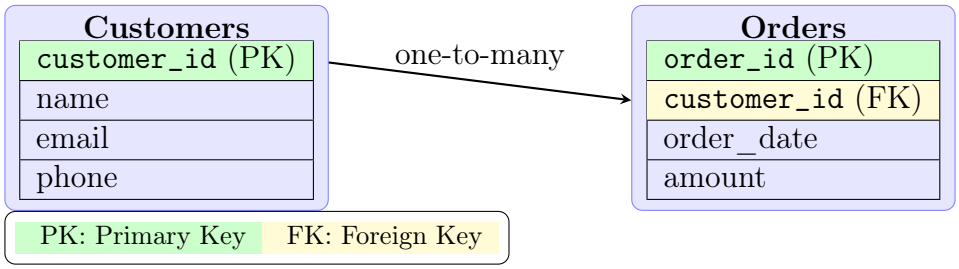

Figure 4: A simple relational database with two linked tables. The "Customers" table (left) stores user data with a unique "customer_id" (green, Primary Key). The "Orders" table (right) records purchases, using "customer_id" (yellow, Foreign Key) to reference the customer who placed each order. The arrow shows a one-to-many relationship: one customer can have multiple orders. Source: Own illustration

To manage relational databases effectively, specialized software is required. This role is fulfilled by a **relational database management system** (RDBMS), which enables users to manage and manipulate relational databases. RDBMS platforms vary in scale and compatibility: some are designed for specific systems, such as personal computers,

while others are increasingly developed to function across more diverse environments and networked platforms. These systems typically rely on **Structured Query Language** (SQL), an industry-standard language used to create and manage database structures and retrieve specific information from relational databases (Taylor, 2003).

While various SQL-based systems exist, the **sqlite3** module in Python offers an especially convenient solution for the workflow presented in this thesis. It allows for integration of a relational database directly within the Python environment, without the need for a separate server (Python Software Foundation, 2025b). The underlying database system, **SQLite**, stores data locally on disk within a database file. This simplicity has helped establish SQLite as one of the most widely used database engines in the world (SQLite, 2025b).

Although relational databases are designed to manage relationships between multiple tables, they can still offer significant advantages even when used with a single table, like in the case of this thesis. As discussed in this subsection, they can enforce constraints and data types (domains), greatly reducing the risk of misformatted or missing data. Another important feature of database systems like SQLite is support for **atomic commits**. This ensures that all changes made during a transaction are applied as a single, indivisible unit, so if any part of the operation fails, none of the changes are saved. This protects against partial updates, which could otherwise compromise data integrity (SQLite, 2025a).

3.3.2 Serialization Formats

Before next discussing serialization formats, it is helpful to first clarify what **data serialization** means. It refers to the process of transforming an object's state into a structured format that can be stored or transmitted. This representation captures all necessary information to recreate the original object later through **deserialization**. The result of serialization is often referred to as an **archive**, and the reconstructed object is functionally equivalent to the original (Grochowski et al., 2019).

Data can be serialized into a variety of formats, such as

- **XML** (eXtensible Markup Language), a text-based format commonly used in web services and valued for its readability across platforms (Hericko et al., 2003).
- **YAML** (YAML Ain't Markup Language), a lightweight, human-readable format

designed for easy data editing with a structure that also supports more complex data (Eriksson and Hallberg, 2011).

- **Binary formats** such as Google *Protocol Buffer*, which generally yield high performance, albeit at the cost of human readability (Aihkisalo and Paaso, 2011).

Another possible format is **JSON** (JavaScript Object Notation), mentioned before as the format that is often returned by REST APIs. JSON serves as a data interchange format that is both human-readable and machine-parsable, while improving performance over XML (Nurseitov et al., 2009).

It was selected as the data serialization format for this thesis due to its superior processing efficiency compared to XML, while maintaining human readability, which binary formats lack. Although YAML also offers readability, JSON was ultimately preferred for its wider adoption (Eriksson and Hallberg, 2011) and better integration with Python, as it is natively supported through the built-in **json** module (Python Software Foundation, 2025a).

3.4 Data Analysis - NLP

This section provides the theoretical framework for the **natural language processing** (NLP) techniques used in this thesis to analyze the collected data. Specifically, it outlines the foundations and methodological principles of **sentiment analysis**, **named entity recognition** (NER), and **topic modeling**.

3.4.1 Sentiment Analysis

Sentiment analysis is a natural language processing technique used to identify and interpret subjective information within textual data, enabling the extraction of individuals' attitudes towards particular topics, entities, products, and so on (Wankhade et al., 2022).

Within the context of data collected from social media platforms, there are a variety of sentiment analysis approaches available. As outlined by Sharma et al. (2020), these are:

- **Supervised Learning**: Utilizes annotated datasets to train classification algorithms such as support vector machines and neural networks, enabling the automated identification of sentiment polarity in unseen data.

- **Unsupervised Learning**: Applies clustering or pattern discovery techniques (e.g., spectral clustering) to infer sentiment structure from unlabeled text, often used when labeled data is scarce or unavailable.

- **Lexicon-Based Approaches**: Employ sentiment lexicons (precompiled lists of words with associated polarity scores) to determine sentiment, either via dictionary-based lookups or corpus-driven methods.

Among the lexicon-based methods, particular tools have been designed to capture the unique characteristics of informal, user-generated content on social media platforms. One such tool is **VADER** (Valence Aware Dictionary for sEntiment Reasoning), a rule-based sentiment analysis model developed specifically for short, social-text formats such as those found on *Reddit*, *Twitter/X*, or *Facebook*. Unlike traditional lexicons that often struggle with informal language or intensity cues, VADER was explicitly constructed and validated to handle such things effectively. To enhance its performance, VADER incorporates a set of rules designed to account for syntactical features that influence sentiment intensity. These include punctuation (e.g., exclamation marks), capitalization (e.g., ALL CAPS for emphasis), degree modifiers (e.g., "very"), negation handling, and contrastive conjunctions (e.g., "but"). This allows VADER to assess not just the polarity (positive or negative), but also the strength of sentiment. It has demonstrated strong performance in analyzing social media text, achieving accuracy levels comparable to or even surpassing those of human raters (Hutto and Gilbert, 2014). This effectiveness, especially in handling informal online language, makes it well-suited for sentiment analysis on platforms like *Reddit*, which is why is was chosen as part of the proposed workflow for this thesis.

While VADER is effective in detecting sentiment in informal, social-media-style language, it is not specifically tuned to financial vocabulary and context. To address this limitation, this thesis also incorporates **FinBERT**, a transformer-based language model adapted for sentiment analysis in the financial domain. FinBERT is a variant of **BERT** (Bidirectional Encoder Representations from Transformers) that has been further trained on financial texts, enabling it to better understand the language often used in economic discussions. Evaluations have shown that FinBERT outperforms both conventional machine learning and other pre-trained models on financial sentiment benchmarks, achieving

high accuracy scores (Araci, 2019). This makes FinBERT a particularly useful addition for analyzing sentiment around topics like the Swiss economy as discussed on *Reddit*.

While *Reddit* posts tend to be more informal than the texts on which FinBERT was trained, it remains plausible that many discussions within economy-oriented posts still incorporate financial language and market-related terminology. Furthermore, transformer-based models such as BERT (and by extension FinBERT) have demonstrated robustness to slight domain shifts (Wankmüller, 2024). Therefore, despite the difference in tone, FinBERT remains a suitable tool for extracting sentiment from *Reddit* posts discussing the Swiss economy. However, as Scherrmann (2023) notes, FinBERT, like all BERT-based models, has an internal token limit of 512, a constraint that must be considered during implementation.

By integrating and comparing both tools, this study aims to leverage their respective strengths: VADER's sensitivity to tone and style, and FinBERT's contextual understanding of finance-specific content. This approach supports a more nuanced and accurate analysis of economic sentiment as it is expressed in both everyday language and financial discourse on *Reddit*.

While this provides a robust framework for sentiment classification, the quality and structure of the input data remain critical factors in determining the success of the analysis. Therefore, appropriate preprocessing steps must be taken to prepare the textual data. This has been shown to improve both the accuracy and efficiency of sentiment analysis tasks (Nikhila Kanigiri et al., 2024). A variety of preprocessing techniques can be applied depending on the context and data source. As stated by Pradha et al. (2019), common methods include:

- **Lowercasing**: Converting all characters to lowercase to avoid treating the same word in different cases (e.g., "Economy" vs. "economy") as distinct entries.
- **URL Removal**: Eliminating web links (e.g., those beginning with "http" or "https") that do not contribute meaningful sentiment.
- **Stemming**: Reducing words to their root form to unify variations (e.g., "invest", "investing", "investment") and streamline analysis.

However, this does not imply that all preprocessing techniques should be universally applied to sentiment analysis tasks. For instance, Bao et al. (2014) demonstrated that, in

the context of a *Twitter/X* dataset, retaining URLs can enhance sentiment classification performance, whereas applying stemming may actually reduce accuracy.

Together, the selection of suitable models and the application of appropriate preprocessing techniques form the foundation for reliable sentiment analysis, enabling a meaningful interpretation of public opinion as expressed on *Reddit*.

3.4.2 Named Entity Recognition (NER)

While sentiment analysis reveals the emotional tone of discussions, it does not provide information about what or who those sentiments are directed towards. To address this, another step in the analytical pipeline is **named entity recognition** (NER), which refers to the process of locating and classifying specific pieces of text that represent real-world entities (such as people, organizations, places, dates, or monetary values) into predefined categories (Marrero et al., 2013).

Similar to sentiment analysis, different approaches for NER have been explored over the years. As Mansouri et al. (2008) point out, these include:

- **Rule-based Approaches**: Rely on handcrafted linguistic rules and patterns combined with curated resources like dictionaries or gazetteers. They often perform well in narrowly defined domains but tend to be less adaptable to new topics or languages due to their fixed rule sets.

- **Machine Learning-Based Approaches**: Treat NER as a classification problem and use statistical or algorithmic models to learn patterns from annotated data. Common techniques include support vector machines and decision trees. While **supervised methods** require large labeled datasets to perform well, **unsupervised approaches** use unannotated data and are less common but potentially more flexible across domains.

- **Hybrid Approaches**: Combine rule-based systems with machine learning techniques to leverage the strengths of both. For example, handcrafted rules might improve precision in known cases, while machine learning models help generalize to unseen examples. However, the limitations of rule-based components, such as reduced portability, may still affect overall adaptability.

In recent years, deep learning has emerged as a new standard in NER, offering substantial improvements over the previously discussed approaches. Unlike feature-based models that rely on extensive manual engineering, deep learning methods automatically learn useful representations from raw input via layered neural architectures. This ability to extract complex features without manual intervention significantly enhances adaptability across domains and languages. These strengths have enabled deep learning models to consistently demonstrate remarkable performance scores. As a result, deep learning-based models are increasingly used in modern NER applications (Li et al., 2022).

Given the proven advantages of deep learning-based methods for NER, this thesis adopts the **SpaCy** framework for named entity recognition. SpaCy uses a hybrid model that incorporates deep learning techniques (specifically, convolutional neural networks) on top of statistical components (Shelar et al., 2020).

Although SpaCy demonstrates strong practical performance, Schmitt et al. (2019) found that its hybrid approach is outperformed by other natural language processing tools, such as **StanfordNLP**. Since version 3.0 however, SpaCy has incorporated transformer-based pipelines, significantly enhancing its accuracy and aligning it more closely with state-of-the-art NLP models (Explosion AI, 2025b), making it a sound choice for the task of NER within *Reddit* posts.

This NER component acts in a supporting role to sentiment analysis by helping identify which entities are being discussed in relation to the Swiss economy. While it is not the central focus of the study, it helps to provide further contex.

This thesis relies on SpaCy's general-purpose NER model, but it could be adapted or retrained to better capture domain-specific entities relevant to economic discourse. However, given time constraints and the broader scope of this project, where NER serves as a supporting component only, such customization falls outside the scope of this study. Instead, the emphasis remains on sentiment analysis as the core analytical objective.

Similarly to sentiment analysis, effective preprocessing of input text is critical for achieving optimal NER performance. Recent research indicates that certain preprocessing steps, such as lowercasing and lemmatization, can significantly influence the accuracy of NER systems. For example, converting all text to lowercase may reduce it by obscuring distinctions between entities, such as "Apple" (the company) and "apple" (the fruit). Similarly, lemmatization, which reduces words to their base form, can also negatively

affect performance. In contrast, proper tokenization and accurate recognition of multi-word expressions (e.g., "White House") have been shown to improve NER outcomes (Chai, 2023). This highlights the need for careful preprocessing choices when applying NER.

In conclusion, named entity recognition, like sentiment analysis, is highly sensitive to the chosen approach and relies heavily on appropriate text preprocessing to achieve optimal performance.

3.4.3 Topic Modeling

With sentiment analysis capturing the emotional tone of *Reddit* posts and named entity recognition helping to identify specific entities being discussed, **topic modeling** is a further step of the analysis. It is a statistical method used to uncover hidden thematic structures in large collections of unstructured text by trying to identify patterns in word usage that point to underlying topics (Vayansky and Kumar, 2020), potentially providing further data insights.

As with sentiment analysis and NER, topic modeling has developed through several methodological approaches. As explained by Abdelrazek et al. (2023), the most prominent are:

- **Probabilistic Models**: Assume a generative process where documents are produced by a mix of latent topics, and each topic is a distribution over words. They are intuitive, interpretable, and modular, with Latent Dirichlet Allocation (LDA) being the most widely used example.
- **Neural Models**: Built on deep learning frameworks, neural topic models replace traditional inference with optimization. They often use embeddings or transformers to capture context, though they may sacrifice interpretability.
- **Algebraic Models**: Rely on matrix decomposition techniques like Non-negative Matrix Factorization (NMF) or Latent Semantic Indexing (LSI) to uncover hidden structure in term-document matrices. They are computationally efficient but lack a statistical foundation.

Across the various performance metrics evaluated by Abdelrazek et al. (2023), neural models frequently rank among the top-performing approaches. The authors further note

that these models may even outperform more traditional techniques in terms of stability when applied to corpora composed of short texts, stating that their results "[...] could suggest that neural topic models outperform classical topic models in stability scores for datasets with short documents [...]" (Abdelrazek et al., 2023, p. 12). This is particularly relevant for this research, as *Reddit* posts (similar to much social media content) tend to be brief, making the strengths of neural topic modeling approaches especially applicable.

Among neural topic modeling techniques, **BERTopic** stands out for its consistent performance across different text types. It uses an embedding-and-clustering approach by employing pre-trained transformer-based models to create document embeddings, which are then clustered into groups. To generate interpretable topic labels, BERTopic applies a variant of **TF-IDF** (Term Frequency - Inverse Document Frequency) based on classes, which identifies the most representative terms for each cluster. This separation of embedding, clustering, and topic representation makes BERTopic highly modular. For example, the language model used for embeddings could be swapped to potentially improve performance in the future. In comparative evaluations, BERTopic performs competitively, even on short and informal text such as tweets. Its performance remains stable regardless of the specific language model used for embeddings, which makes it both accessible if computing power is limited and future-proof. Additionally, it supports **dynamic topic modeling**, which enables the analysis of how topics evolve over time (Grootendorst, 2022). Given the temporal nature of this study's research question, this, together with its previously mentioned advantages, results in making BERTopic the most suitable choice for this research.

Similarly to sentiment analysis and NER, text data must undergo thorough preprocessing before applying topic modeling techniques, in order to generate optimal results. Wachirapong (2023) emphasizes that decisions made during preprocessing can significantly affect the resulting topics produced, while underlining that there is no universal approach. Rather, preprocessing choices should align with the characteristics of the text and the research goals. For *Reddit* posts, which are often short and informal, proper preprocessing helps to ensure that the BERTopic model can extract coherent and interpretable topics from the data.

During the evaluation of BERTopic, its authors applied standard preprocessing steps such as lowercasing, stopword and punctuation removal, as well as lemmatization, par-

ticularly for longer and formal datasets. For shorter, informal texts like tweets, they opted for more minimal preprocessing (Grootendorst, 2022). This emphasizes the results of Wachirapong (2023): both level and type of preprocessing should be adapted to the nature of the dataset and analysis goals.

3.5 Choice of Programming Language - Python

As was briefly mentioned throughout this thesis so far, **Python** was chosen as the primary programming language for this study due to its strong ecosystem of libraries for natural language processing and data visualization. Widely used in both academic and industry settings (Hunt, 2023, p. 15), Python supports tools such as **VADER** and **FinBERT** for sentiment analysis (Python Software Foundation, 2025c; Araci and Genc, 2022), **SpaCy** for named entity recognition (Explosion AI, 2025a), and **BERTopic** for topic modeling (Grootendorst, 2024), all of which were established as adequate choices for this research in the previous sections.

In addition to supporting the analytical components of the pipeline, Python also facilitates data acquisition and storage. As mentioned earlier in this chapter, the **PRAW** and **AsyncPRAW** libraries can interact with the *Reddit* API, while the **sqlite3** and **json** modules provide solutions for storing and managing the collected data. Furthermore, libraries such as **Seaborn** and **Matplotlib** offer tools to visualize results, enabling better interpretation and communication of findings (Waskom, 2021; Hunter, 2007).

3.6 Measuring Economic Performance in Switzerland

To evaluate whether public sentiment on *Reddit* reflects and possibly predicts economic developments in Switzerland, it is essential to identify indicators that capture key dimensions of the country's economic performance. Since this study aims to explore potential correlations between online discourse and real-world conditions, such indicators serve as valuable reference points for interpreting sentiment trends. The following section outlines four economic indicators that offer complementary perspectives on Switzerland's economy and are particularly well suited for comparison with the sentiment data collected in this study:

- **Gross Domestic Product** (GDP) serves as the most comprehensive indicator

of economic performance. Published quarterly by the Swiss State Secretariat for Economic Affairs (SECO) (2025c), it reflects the overall output of goods and services and is commonly used to assess economic growth or contraction (Callen, 2008).

- The Swiss **Consumer Price Index** (CPI) measures inflation and changes in the cost of living. Released monthly by the Swiss Federal Statistical Office (FSO) (2025), the CPI can reasonably be assumed to influence public sentiment, as rising prices often affect purchasing power and household budgets.

- The **Swiss Market Index** (SMI) tracks the performance of 20 major publicly traded Swiss companies (SIX Group, 2025), making it a useful proxy for investor confidence and market expectations. Although more volatile than GDP or CPI, the SMI still provides insight into financial sentiment, which may be echoed in online discussions.

- Lastly, the **Consumer Sentiment Index**, published monthly by the State Secretariat for Economic Affairs (SECO) (2025a), offers a survey-based measure of how Swiss individuals perceive the current economic climate and future expectations. As a direct reflection of public opinion, it provides a benchmark for comparing formal survey-based sentiment with informal, user-generated discourse on *Reddit*.

Since the data in this study is measured on an annual basis, all economic indicators will be sourced as or transformed to yearly values to ensure temporal alignment. While this reduces granularity, it ensures consistency when comparing data trends to key economic indicators. Together, the theoretical foundations outlined in this chapter form the basis for the analytical approach implemented in the following chapter.

4 Methodology

4.1 Introduction and General Overview

This chapter outlines the methodological framework used to investigate the relationship between sentiment on *Reddit* and economic developments in Switzerland. Building on the theoretical foundations established in the previous chapter, it details the data collection process, preprocessing steps, and the implementation of sentiment analysis, named entity

recognition, and topic modeling. It also describes how the results were evaluated against selected economic indicators to explore potential relationships and addresses important limitations. The aim is to provide a transparent and reproducible account of the analytical pipeline developed for this study. Figure 5 (seen below) offers a visual summary of the workflow:

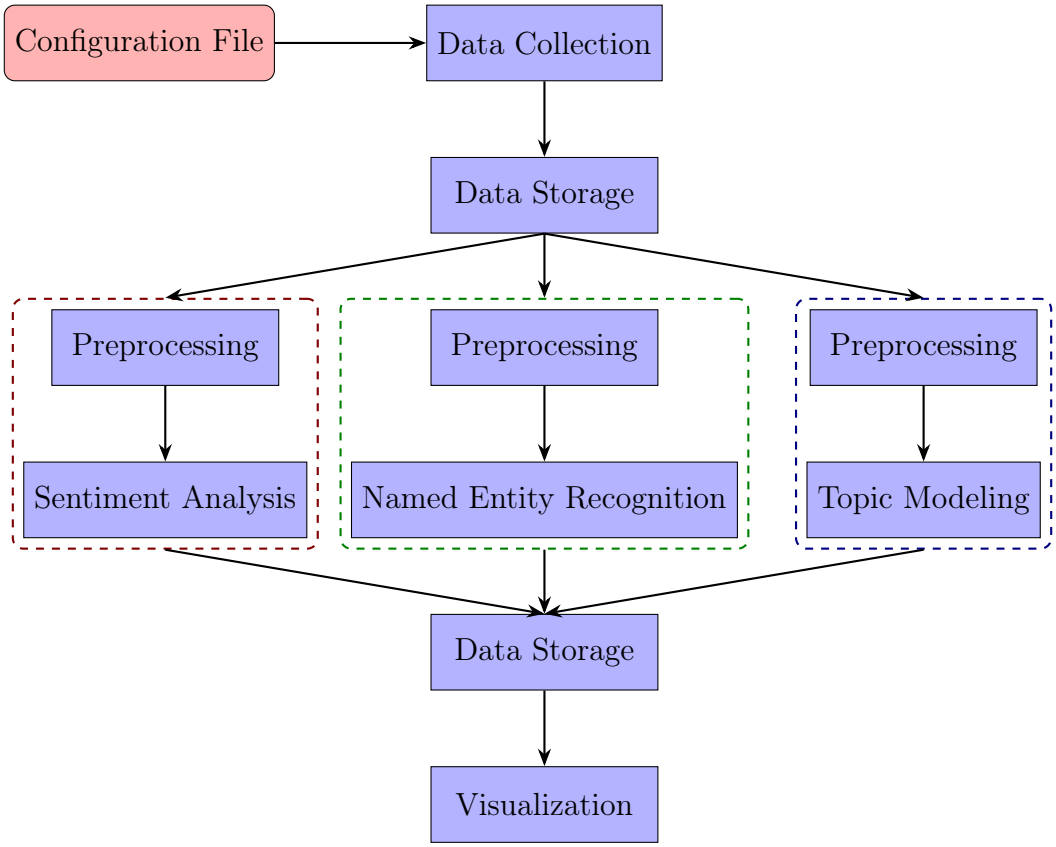

Figure 5: The collection and analysis pipeline developed during the course of this thesis. Source: Own illustration

The full implementation of the analytical pipeline is available on GitHub. For security and privacy reasons, sensitive credentials such as *Reddit* API login data have been removed.

The following sections describe each component of the pipeline in the order shown above.

4.2 The Configuration File

The **configuration file**, stored in JSON format under the name `config.json`, serves multiple purposes within the data collection step. Most importantly, it separates sen-

sitive information such as *Reddit* API login credentials from the main Python scripts, enhancing security. In addition, it allows data collection parameters (such as subreddit name, query terms or sorting methods) to be easily configured without the need to modify the underlying code.

The necessary login credentials are stored under the keys `client_id`, `client_secret`, and `user_agent`, which are required for authenticating with the *Reddit* API. The remaining keys in the `config.json` file are used to control the data collection process. Each is described below:

- `limit`: Specifies the maximum number of items to retrieve per API query. As explained in Section 3.2.2, *Reddit*'s API limits this to a maximum of 1,000 items per query.

- `subreddit`: Defines the specific subreddit to search within. For this study, this is set to **"Switzerland"**.

- `query_terms`: Contains a list of keywords used to filter posts relevant to the Swiss economy. These terms were chosen to maximize the likelihood of retrieving economically relevant content. During testing, it was observed that including too many keywords in a query, particularly those that were overly specific or used very infrequently in everyday language, could significantly reduce the number of results returned by the *Reddit* API. To address this, the final list was carefully curated through iterative testing, retaining only those terms that consistently retrieved a sufficient volume of posts. The specific list used is: **["economy", "inflation", "GDP", "finance", "banking", "SNB", "UBS", "Credit Suisse", "employment", "unemployment", "taxation", "cost of living", "Swiss Franc", "recession", "wages", "salary", "job market", "rent", "real estate", "housing market", "interest rates", "exchange rate", "forex", "SMI", "Swiss companies", "Swiss banking"]**.

- `sorting_methods`: Determines which sorting options the *Reddit* API should apply when searching for posts. To increase coverage and diversity, all major sorting methods are included: **["relevance", "hot", "new", "top", "comments"]**.

- `sample_size`: Specifies the number of posts to randomly sample per year from the total pool of retrieved posts for that year. A size of **100** was selected to balance

processing efficiency with data representativeness. This parameter is only used if the retrieved posts for a given year exceed `sample_size`.

- `start_year`: Indicates the earliest year from which to collect posts. This was set to **2008**, the year the subreddit r/Switzerland was created, to ensure a complete historical scope.

- `reset_db`: A Boolean value that controls whether the database is cleared before each execution of the data collection script. For a clean start, this is set to **true**.

These configuration settings are then used to guide the automated data collection process described in the following section.

4.3 Data Collection - PRAW / AsyncPRAW

This process happens within a Python script named `Data_Collection.py`. It uses an asynchronous architecture based on the **AsyncPRAW** library to query *Reddit*'s API and retrieve posts from the subreddit r/Switzerland across a defined time range. This asynchronous design significantly improves efficiency by allowing multiple years of data to be collected in parallel, rather than sequentially. The script searches for posts using the economy-related keywords and multiple sorting methods defined in the configuration file to increase coverage and reduce bias introduced by *Reddit*'s internal ranking algorithms. Posts retrieved across different sort methods may include duplicates, but only unique entries are stored in the database.

To support year-over-year analysis, posts are assigned to specific calendar years based on their creation timestamps by employing the **datetime** library and grouped accordingly. The script performs automatic language detection using the **langdetect** library, retaining only posts where both the title and body text are identified as English and where the body text is not empty. To prevent duplicate entries resulting from overlapping query results, each post ID is checked before the post is saved for later analysis. When more than the configured number of unique and valid posts (given by the `sample_size` parameter in the configuration file) are available for a given year, a random sample is drawn using a fixed seed to ensure reproducibility.

4.4 Data Storage - SQLite & JSON

The collected *Reddit* data and subsequent analysis results are stored using two main formats. The original post metadata, such as titles, text content, timestamps, as well as sentiment analysis results placeholders and final results, are stored in a local **SQLite** database using the **aiosqlite** library, an asynchronous variant of Python's built-in **sqlite3** module. This database is named `swiss_economy_reddit.db`. As described in Section 3.3.1, SQLite allows for the enforcement of constraints and promotes data integrity, even though this study utilizes only a single table, `reddit_posts`. Figure 6 (below) illustrates the database schema in greater detail:

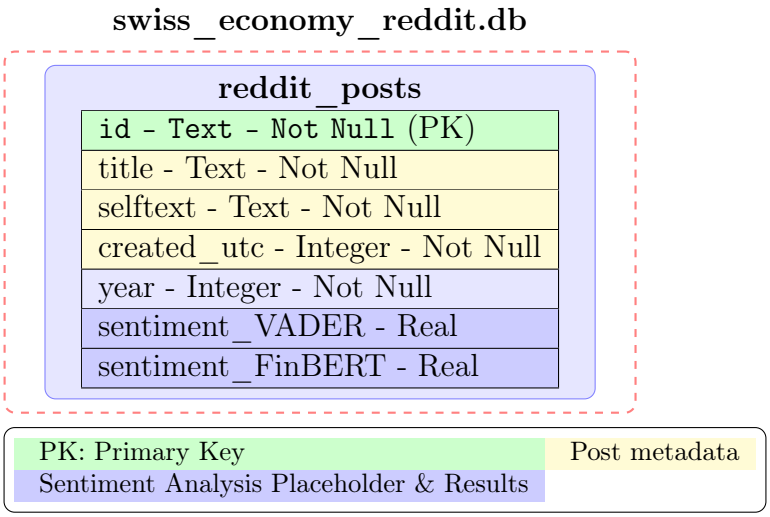

Figure 6: The SQLite database used by this thesis, with data types and enforced constraints. Each post's unique ID is used as primary key, ensuring no duplicate posts. Further, the database contains post title, text, timestamp of creation, year of creation (assigned by the collection script) as well as two placeholders for the results of the later sentiment analysis. Source: Own illustration

In contrast, the outputs of later analysis steps, specifically named entity recognition and topic modeling, are saved in **JSON** format, similar to the configuration file. For NER, the saved results include annotated entities per post, as well as aggregated frequency counts and structured entity lists, both at the yearly and overall level. For topic modeling, the JSON output includes topic assignments and relevance probabilities for each post, as well as summaries of the identified topics (e.g., keywords and sizes), grouped by individual years and an aggregated category for all years. As already outlined in Section 3.3.2, the JSON format was chosen due to its human readability and seamless Python integration.

However, in the case of topic modeling, storing results in JSON alone is insufficient for

enabling more advanced visualizations and interactions. BERTopic requires access to the full trained model to generate interactive plots or explore topic embeddings. Therefore, the complete BERTopic models were saved separately alongside the JSON file to support such functionality. As these are binary files, they are not human-readable.

With storage formats established, the next step involves preparing the collected and stored data for analysis. This requires a series of preprocessing operations designed to enhance linguistic consistency and reduce noise in the text.

4.5 Preprocessing

As discussed previously in Sections 3.4.1, 3.4.2 & 3.4.3, careful data preprocessing is essential for optimizing analytical performance. However, it has also been shown that there is no one-size-fits-all approach; different analysis steps benefit from different preprocessing techniques to enhance their respective outcomes. The following section outlines the preprocessing steps applied individually for sentiment analysis, NER, and topic modeling. These procedures are not implemented in a separate script, but integrated directly into each analysis script and executed prior to the respective processing stages, as is indicated in Figure 5.

- **Sentiment Analysis**: In line with the research discussed in 3.4.1, a minimal approach (only lowercasing) was initially considered as a preprocessing step for sentiment analysis. However, as discussed previously, VADER relies on capitalization cues as part of its rule set. Consequently, this normalization step was omitted to preserve these features. In contrast, FinBERT does not depend on such surface-level formatting. Therefore, all text was lowercased prior to analysis when using FinBERT.

- **Named Entity Recognition**: No explicit preprocessing was applied prior to named entity recognition, as the transformer-based SpaCy model used in this study includes its own tokenization, case handling, and multi-word expression recognition, which were presented in Section 3.4.2 as crucial components for NER performance. These preprocessing steps are integrated directly into the model, reducing the need for manual intervention. As a result, the raw text was passed directly to the model.

- **Topic Modeling**: Given the short and informal nature of *Reddit* posts, which are

comparable to the tweet data used by Grootendorst (2022) to evaluate BERTopic, extensive normalization was intentionally avoided at first. Specifically, only lowercasing was performed to standardize input. This approach aligns with the original evaluation of BERTopic on short-text datasets. However, preliminary analysis revealed that this minimal preprocessing approach resulted in suboptimal topic extraction, with many topics dominated by e.g. stop words rather than meaningful semantic content. To address this, additional cleaning steps were implemented prior to BERTopic processing:

- **Basic normalization**: URLs, punctuation, and numerical values were removed while preserving alphanumeric characters and standardizing whitespace.
- **Selective stop word removal**: Common English stop words were filtered while retaining negation terms (e.g., "not", "no") to preserve critical semantic relationships.
- **Length-based filtering**: Posts with fewer than four words were excluded to eliminate non-substantive content.

Additionally, for all three analysis steps, the title and body text of each post were merged together to form a single input sequence.

Following the tailored preprocessing for each analytical component, the main analytical step is sentiment analysis using VADER and FinBERT.

4.6 Sentiment Analysis - VADER & FinBERT

To accomplish this, the script `Sentiment_Analysis.py` was developed. It analyzes each post stored in the SQLite database using both VADER (via the **nltk** library) and FinBERT (via the **transformers** library), subsequently writing the resulting sentiment scores back into the database.

As described in the preceding section, each post's title and body text are concatenated into a single input string prior to analysis. VADER operates on the original, unaltered text in order to utilize formatting features such as capitalization, while FinBERT receives a lowercased version of the input.

VADER returns a compound sentiment score ranging from -1 (most negative) to +1 (most positive), reflecting the overall polarity of the post and can be fed input directly.

In contrast, FinBERT requires more careful input handling due to its maximum token limit of 512, as outlined in Section 3.4.1. To prevent truncation of longer texts, the script implements a chunking strategy, in which each input string is tokenized and split into overlapping segments of 510 tokens. An overlap of 50 tokens is used between chunks to preserve context across segments. This approach ensures that, when FinBERT automatically adds the special `[CLS]` and `[SEP]` tokens, the total input length remains within the allowed 512-token limit.

Each resulting chunk is passed to FinBERT, which returns a predicted sentiment label (positive, negative, or neutral) along with an associated confidence score. These outputs are then mapped to numeric sentiment values: the score is assigned as positive, negative, or zero depending on the predicted class. The final sentiment value for a post is computed by averaging all chunk-level scores, producing a continuous sentiment value centered around zero.

4.7 Named Entity Recognition - SpaCy

Another part of the analytical pipeline is NER, conducted using the script `Named_Entity_Recognition.py`. This script applies SpaCy's transformer-based English model (`en_core_web_trf`) to each post stored in the SQLite database and writes the results to a structured JSON file, named `ner_results.json`. As with sentiment analysis, the title and text of each post are combined to form a single input before being passed to the model.

For processing, the posts are temporarily loaded into a **pandas** DataFrame. NER is then performed using SpaCy's `pipe` method with batching enabled. To optimize runtime, non-essential components of SpaCy's NLP pipeline, namely part-of-speech tagging, dependency parsing, and lemmatization, are disabled. Named entities are extracted as tuples containing the entity text and its corresponding label (e.g., `("UBS", "ORG")`), and the resulting list is stored in a new DataFrame column. The recognized entities are then aggregated to enable both overall and time-specific analysis, with results written to the previously mentioned JSON file for later use.

This approach allows for both general and year-specific insights into the types of entities mentioned in Swiss economic discourse on *Reddit*.

4.8 Topic Modeling - BERTopic

In addition to sentiment analysis and NER, the pipeline involves topic modeling using BERTopic, implemented in the script `Topic_Modeling.py`. This script applies BERTopic to analyze the collected *Reddit* posts using a transformer-based embedding model and stores the results in a structured JSON file (`bertopic_results.json`). To ensure compatibility with data structures such as **numpy** arrays and **pandas** DataFrames, a custom serialization function was developed. As discussed in Section 4.4, the full BERTopic model for each year is saved separately to disk to enable later visualization and exploration. Again, the title and text of each post are joined to form a single string prior to analysis.

The script retrieves all posts from the database and groups them by calendar year. As noted in Section 4.5, all input text is normalized (URL / punctuation / number removal and whitespace standardization using regular expressions) and lowercased prior to processing, while common stop words are filtered via SpaCy's English stopword list (retaining some negations). Additionally, posts containing fewer than four words are filtered out to minimize noise. The remaining posts are embedded using the `paraphrase-MiniLM-L6-v2` model from the **sentence_transformers** library, which converts text into numerical vectors. BERTopic is configured with `calculate_probabilities=True` to generate soft cluster assignments, allowing for a more nuanced understanding of topic relevance per document. Additionally, the number of topics is not fixed manually; instead, the model is allowed to determine the optimal number of topics automatically (`nr_topics=None`). This approach helps balance interpretability and granularity by allowing the topic structure to adapt to the size and complexity of each year's dataset. These embeddings are then clustered using BERTopic's built-in pipeline, which combines **UMAP** for dimensionality reduction and **HDBSCAN** for topic detection.

For each year with at least ten valid posts, as well as for the aggregated corpus of all years, BERTopic is applied independently. The threshold of ten posts was chosen because the underlying components require a minimum number of input samples to generate meaningful and stable topic structures. Posts classified as outliers (assigned to topic -1) are excluded from the final output.

This procedure supports both yearly analysis and an aggregated overview of topic trends within the collected posts.

4.9 Visualization of Results

Since the pipeline stores its outputs in non-visual formats (SQLite, JSON, and binary files), visualizing these results is a critical step for interpretation. This is accomplished using the script `Visualization.py`, which consolidates all visualization logic into a single script. A key design goal was to ensure flexibility: the visualization functions for sentiment analysis, NER and topic modeling are self-contained and only execute if the necessary input file (e.g., `ner_results.json`, `bertopic_results.json`) is present. This makes it possible to run or rerun specific visualizations independently, without re-executing the entire analytical pipeline.

Sentiment analysis results are displayed using **matplotlib** and **seaborn** via a **dual-axis plot**: a bar chart shows the number of analyzed posts per year, while two line plots represent the average VADER and FinBERT sentiment scores over time. In addition, horizontal lines indicate the overall average sentiment for each method across all years. No further transformation or preprocessing of sentiment data occurs at this stage.

Conversely, before visualizing the NER results, a light cleaning step is applied to improve interpretability. This involves filtering out entities that are either likely to introduce noise or do not meaningfully contribute to the analysis. Specifically:

- Common stopwords, as defined by the SpaCy English stopword list, are excluded.
- Entities consisting solely of digits are removed, unless they belong to categories where numerical values are meaningful, such as `DATE`, `TIME`, `MONEY`, `PERCENT`, or `CARDINAL`.
- Very short or unusually long strings (fewer than 2 or more than 30 characters) are also filtered out.
- Only entities that match a simple regular expression (alphanumeric with optional punctuation such as hyphens or apostrophes) are retained.

Cleaned NER results are then visualized using a mix of **plotly**, **matplotlib**, and **wordcloud**:

- **Bar charts** showing the ten most frequent entity types (e.g., `ORG`, `GPE`, `PERSON`).
- **Word clouds** highlighting the 100 most frequently mentioned named entities.

- **Treemaps** providing an interactive breakdown of entity types and their most common instances.

All of the above entity visualizations are available for both individual years and the full dataset. Additionally, a **line chart** depicting temporal trends for the top five entity types is computed, allowing insights into how attention to different entity categories evolves over time.

Similarly to NER, topic modeling results are visualized both annually and across the entire dataset by using **matplotlib** and **BERTopic's built-in tools**:

- A **bar chart** shows the number of documents assigned to each discovered topic.
- If exactly two topics are identified, a **Venn diagram** illustrates shared and unique keywords across them.
- When more than two topics are found and at least five topic embeddings are available, a UMAP-based **cluster map** is generated using BERTopic's built-in interactive visualization tools.
- Additionally, an **interactive bar chart** of top keywords per topic is created using BERTopic's `visualize_barchart()` method.

Safeguards are implemented to ensure that topic visualizations requiring a sufficient number of topics or embeddings are only generated when appropriate, thus preventing runtime errors.

The script organizes visualizations by creating a main `visualizations` directory with subfolders for each analysis type (sentiment analysis, NER, and topic modeling). All generated charts and interactive visualizations are stored in these directories for future reference and analysis.

4.10 Economic Indicators

With all data collected, analyzed, and visualized, the final step of this thesis involves comparing the sentiment results from *Reddit* discussions with established economic indicators for Switzerland. This comparison aims to address the central research question: **Does user-expressed sentiment on *Reddit* reflect real-world economic developments in Switzerland, and can it potentially serve as a predictive signal?**

As discussed in Section 3.6, the economic indicators used in this study are published at a higher temporal resolution. To ensure compatibility with the *Reddit* data, which was aggregated on an annual basis, all key indicators were likewise sourced as or, if necessary, converted to yearly values. The following measures and transformations were chosen:

- **Gross Domestic Product** (GDP): Year-over-year percentage change, as reported by the Federal Statistical Office (FSO) (2024). This reflects whether the overall economic output increased or decreased compared to the previous year. Since the GDP data published by the Federal Statistical Office (FSO) is only available up to 2023, quarterly figures for 2024 and 2025 provided by the State Secretariat for Economic Affairs (SECO) (2025c) were averaged to approximate annual values for those years.

- **Consumer Price Index** (CPI): Average annual inflation rate, sourced from the Federal Statistical Office (FSO) (2025). This indicates the extent of inflation or deflation in consumer prices. As annual data is only available up to 2024, monthly data was averaged for 2025.

- **Swiss Market Index** (SMI): Year-end values were obtained from Statista (2025) and used to trace the overall development of the Swiss stock market over time. For the ongoing year 2025, the end-of-day value on July 31st was sourced from cash.ch (2025).

- **Consumer Sentiment Index**: Long time series survey results (seasonally adjusted) published by the State Secretariat for Economic Affairs (SECO) (2025b) were averaged to obtain yearly mean values, providing an official benchmark for public economic perception.

To explore potential relationships between economic sentiment expressed on *Reddit* and official economic performance metrics, a primarily exploratory and qualitative approach was adopted. Rather than applying formal predictive models, this study focuses on pattern recognition and interpretive comparison. When notable shifts in sentiment were observed, corresponding movements in economic indicators were examined to assess temporal alignment or possible lagged effects. The goal is to surface potential patterns and relationships rather than establish causality.

4.11 Data Limitations and Methodological Considerations

While this thesis presents a structured and reproducible pipeline for analyzing the sentiment, entities and topics within Swiss economic discourse on *Reddit*, several key limitations must be acknowledged. These arise from both the nature of the data and the methodological choices made throughout the study. For the sake of transparency and clarity, the following section categorizes, outlines, and discusses important limitations across data quality and model performance.

4.11.1 Data Limitations

In terms of data quality, several limitations should be noted. First, *Reddit*'s API is somewhat volatile, occasionally returning different results for the same query within a short time frame. This introduces a degree of inconsistency to the data collection process.

Second, while economy-focused keywords were used to improve the relevance of the retrieved posts, their effectiveness is inherently limited: they can only increase the likelihood that a post relates to the economy but cannot guarantee topical precision. Consequently, the dataset may contain noise in the form of off-topic or tangentially related posts.

Third, as mentioned in Section 4.2, the inclusion of too many keywords in a single query was found to sometimes drastically reduce the number of results returned by the API, occasionally lowering the count to zero. To mitigate this, the final keyword list was refined based on observed query performance, ensuring adequate recall without introducing overly restrictive queries. This limitation created a trade-off between recall and precision: including fewer keywords helped ensure results were returned, but also risks excluding relevant economic discussions.

Additionally, *Reddit* content can be edited or deleted by users or moderators after collection, introducing the possibility of missing or outdated information. The demographic makeup of *Reddit* further limits generalizability: as discussed in Section 2.1, the platform's user base skews young, male, and English-speaking. This demographic bias means that the discourse captured in the dataset may not fully represent the broader Swiss population. The study also filters for English-language posts exclusively, thereby excluding multilingual or native-language discussions.

Finally, the pipeline is designed to process textual content only. While this focus simplifies analysis, it excludes important contextual elements found in non-textual content

such as images or linked articles. As a result, some nuances of user communication and discussion dynamics may be lost.

4.11.2 Methodological Considerations

Beyond data constraints, the methodological design of this study introduces certain limitations. While the overall pipeline was built for modularity and reproducibility, specific modeling choices may have influenced the results.

For instance, FinBERT, due to its maximum input length of 512 tokens, required a chunking strategy with overlapping segments to preserve context. Although this approach helps prevent truncation, it may introduce minor redundancy or uneven weighting in the final sentiment scores. VADER, on the other hand, was applied without preprocessing steps at all. This is consistent with its original design, but these choices may affect comparability with other sentiment models that assume more normalized inputs.

Moreover, both models are likely to struggle with detecting sarcasm, irony, or other forms of subtle language common in online discourse. And while FinBERT performs well on financial texts and is somewhat resilient to light domain shifts (as described in Section 3.4.1), it was not trained on social media data, which may reduce its effectiveness when applied to informal *Reddit* content.

In the case of NER, the use of SpaCy's transformer-based model (`en_core_web_trf`) offers high-quality entity extraction but abstracts away some preprocessing logic, limiting fine-grained control. Additionally, reliance on standard entity label schemas (e.g., `ORG`, `GPE`, `PERSON`) may constrain interpretability for more domain-specific entities relevant to Swiss economic discourse.

Topic modeling was conducted using BERTopic, which can exhibit performance sensitivity to small input sizes and parameter variation, particularly in years with few posts, where resulting topic structures may be unstable or less meaningful.

Lastly, while effort was made to present results clearly through visualizations, no manual validation against annotated ground truth was performed. As such, the reliability of model outputs (e.g., sentiment scores or NER tagging) remains unconfirmed, and conclusions drawn from them should be interpreted cautiously.

4.12 Conclusion

This chapter has introduced a modular and reproducible methodological framework for investigating economic sentiment on the subreddit r/Switzerland. By combining lexicon- and transformer-based sentiment analysis with transformer-based named entity recognition and topic modeling, this approach aims to capture the tone, entities, and themes of economic discourse, examining whether *Reddit* sentiment reflects real-world Swiss economic developments and possesses predictive potential.

Key economic indicators (GDP, CPI, SMI, Consumer Sentiment Index) are used to contextualize findings, though the study remains exploratory and qualitative. Limitations inherent to the methodology, such as API constraints and model-specific challenges have been acknowledged. The next chapter presents the results of the pipeline developed in this chapter, with their implications, including comparison to key economic indicators, discussed thereafter.

5 Results

5.1 Data Collection Summary

The data collection script returned a total of 288 posts matching the selected economy-focused keywords. The distribution of posts over time is shown below in Figure 7:

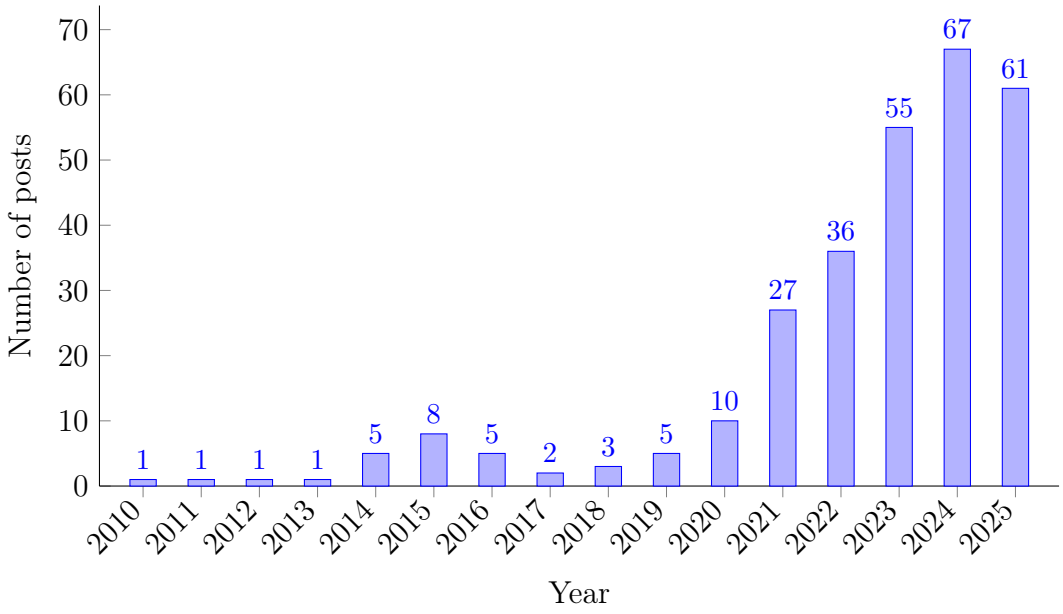

Figure 7: Bar chart showing number of posts collected per year by `Data_Collection.py`. Source: Own illustration

The collected posts spanned 15 years, from 2010 to 2025. From 2010 through 2013, only one post per year was retrieved. Beginning in 2014, the number of posts increased modestly, reaching five posts that year and peaking at eight in 2015, before declining again over the following years.

A more pronounced upward trend begins in 2019, with the number of posts rising from five to ten in 2020, and then increasing sharply year over year. This culminates in a peak of 67 posts in 2024, followed closely by 61 posts in 2025.

5.2 Sentiment Analysis Results

The results of the sentiment analysis are visualized in Figure 8. The plot displays both VADER and FinBERT sentiment trends over time and the number of posts analyzed per year.

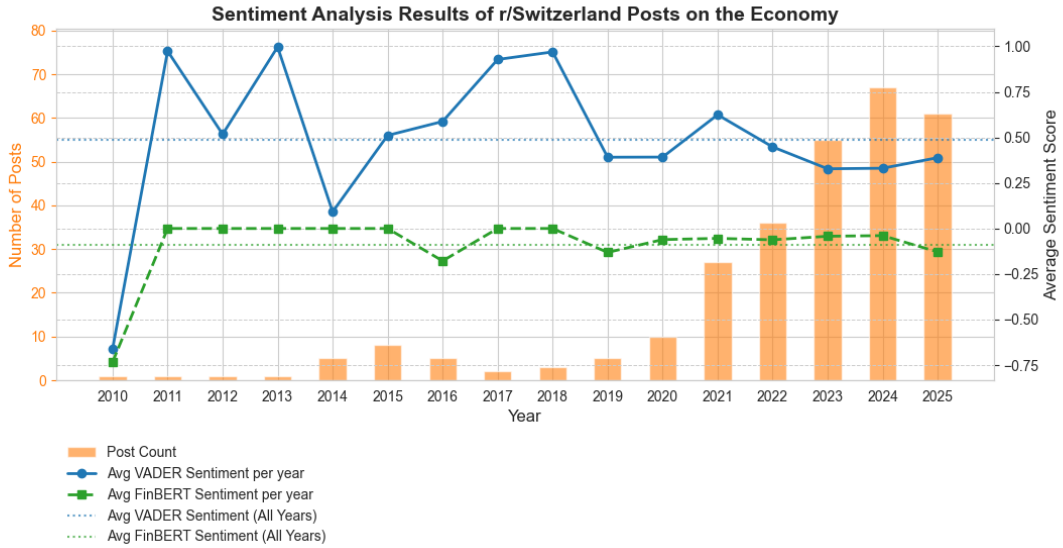

Figure 8: Results of the sentiment analysis performed using `Sentiment_Analysis.py`, showing average yearly and global sentiment scores from VADER and FinBERT alongside the number of analyzed posts between 2010 and 2025. Source: Own illustration

Across all years, VADER consistently yields higher sentiment scores than FinBERT, with the latter remaining closer to neutral or slightly negative. This is also reflected in the overall average sentiment scores: approximately **0.5** for VADER and around **−0.1** for FinBERT.

Both models display fluctuations in yearly sentiment scores, which are especially heavy during earlier years with low post counts. From 2019 onward, as the number of posts

increases, the sentiment trends appear more consistent over time.

Starting from this point, VADER's sentiment score remains unchanged in 2020 before peaking in 2021. From there, it gradually declines until 2023, exhibits no change in 2024, and rises slightly again in 2025. FinBERT's score, on the other hand, increases between 2019 and 2020, then remains relatively stable through 2022, followed by a modest rise. Like VADER, FinBERT shows no change in 2024 but drops noticeably in 2025, creating a contrast in sentiment trajectories for that final year.

5.3 Named Entity Recognition (NER) Results

The results of SpaCy's transformer-based named entity recognition are presented below. As outlined in Section 4.9, bar charts, treemaps, and word clouds were generated for each year as well as for the entire dataset. To enhance clarity and avoid overcrowding, this section focuses solely on the aggregated results. A more detailed discussion of the yearly trends will follow in the next chapter.

To begin, the temporal evolution of entity type frequencies across the dataset is examined. This is shown below in Figure 9:

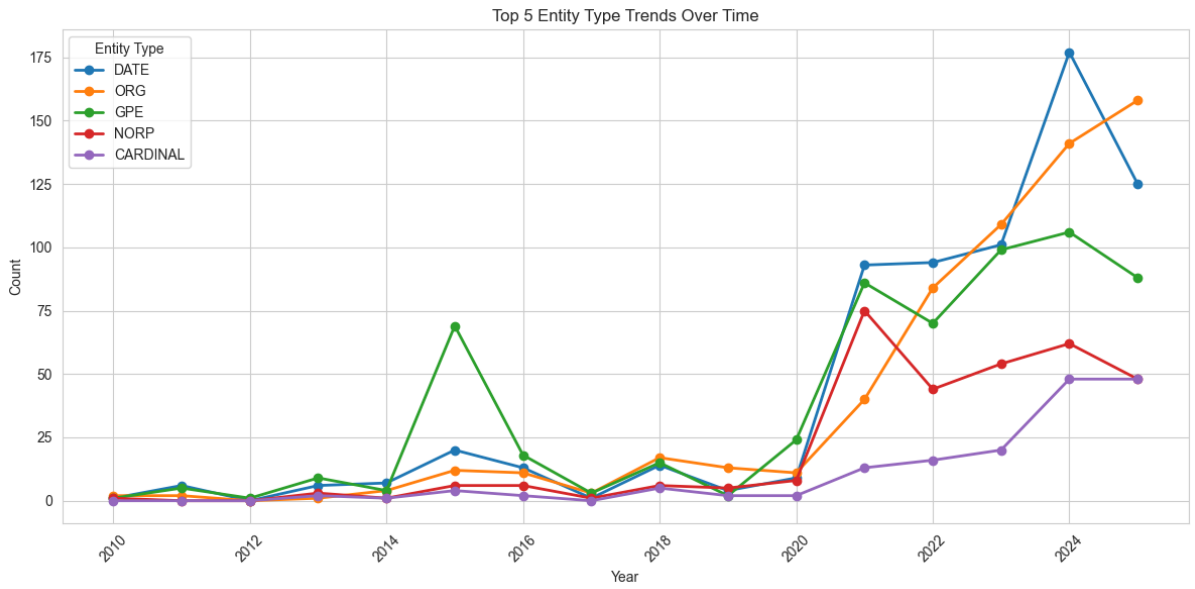

Figure 9: Temporal trends of the top 5 most frequent entity types (`DATE`, `ORG`, `GPE`, `NORP`, `CARDINAL`) from 2010–2023. Entity types are defined as follows: `DATE`: dates/times, `ORG`: organizations, `GPE`: geo-political entities, e.g., countries/cities, `NORP`: nationalities/religious/political groups, `CARDINAL`: numerical values. Source: Own illustration

In the earlier years (2010–2013), the count of recognized entities is consistently low,

aligning with the limited number of collected posts for this period. Between 2014 and 2015, entity recognition rates rise sharply, with geo-political entities (GPE) exhibiting a very pronounced spike. After 2015, the counts decline again, remaining low until 2020. This trend closely mirrors the overall post frequency in the dataset during these years.

From 2020 onward, entity recognition rates increase noticeably. Organizations (ORG) demonstrate consistent growth through 2025, numerical values (CARDINAL) follow a similar but less pronounced trajectory, plateauing between 2024 and 2025.

The remaining three entity types exhibit distinct fluctuation patterns. DATE entities experience a sharp initial increase, stabilize somewhat between 2021 and 2023, then surge again before declining in 2025. GPE shows cyclical variation: increasing (2020-2021), decreasing (2022), rebounding (2023-2024), and falling again in 2025. Nationalities/religious/political groups (NORP) mirror this periodicity, with parallel increases and decreases occurring in the same years as GPE entities, though at lower absolute frequencies.

Following the temporal analysis, Figure 10 presents the distribution of entity types across the complete dataset. The bar chart displays the ten most frequent entity types aggregated over all study years (2010-2025), revealing their relative proportions in the corpus.

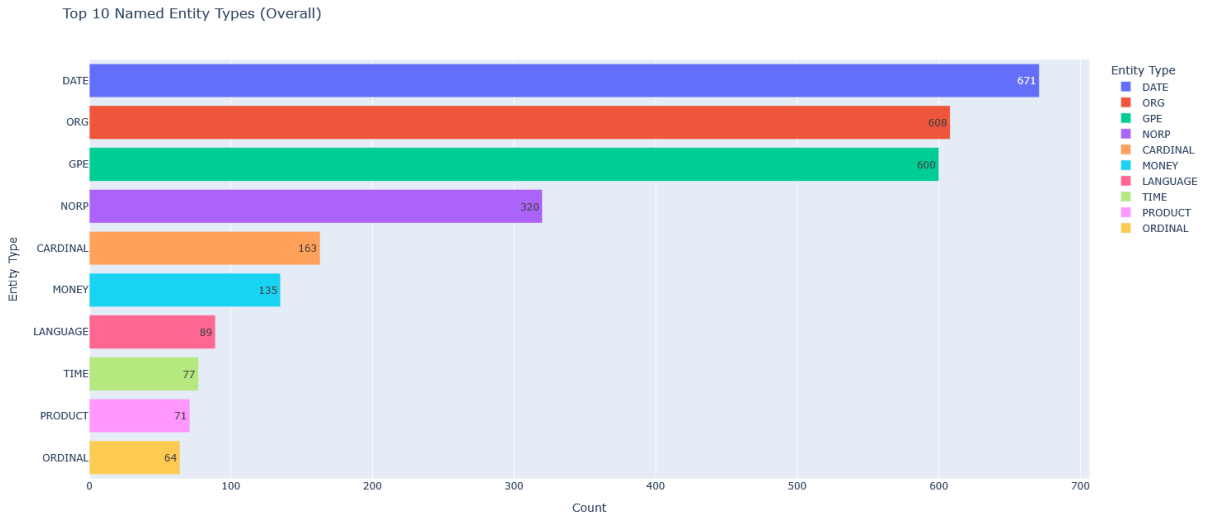

Figure 10: Frequency distribution of the top 10 named entity types across the entire dataset (2010–2025). Entity types are defined as: DATE: dates/times, ORG: organizations, GPE: geo-political entities, NORP: nationalities/religious/political groups, CARDINAL: numerical values, MONEY: monetary values, LANGUAGE: language names, TIME: clock times, PRODUCT: commercial products, ORDINAL: ordinal numbers. Values represent total counts per entity type. Source: Own illustration

DATE entities dominate the distribution with 671 occurrences, followed closely by ORG (608) and GPE (600). A significant frequency drop occurs with NORP (320), which is approximately half as prevalent as the top three types. CARDINAL entities (163 counts) appear at roughly half the frequency of NORP, followed by MONEY (135). The remaining types (LANGUAGE, TIME, PRODUCT, ORDINAL) are all falling below 100 recognized instances each.

Complementing the frequency rankings, a treemap spatially represents dominance relationships among entity categories and their constituent terms, with area sizes corresponding to occurrence frequencies.

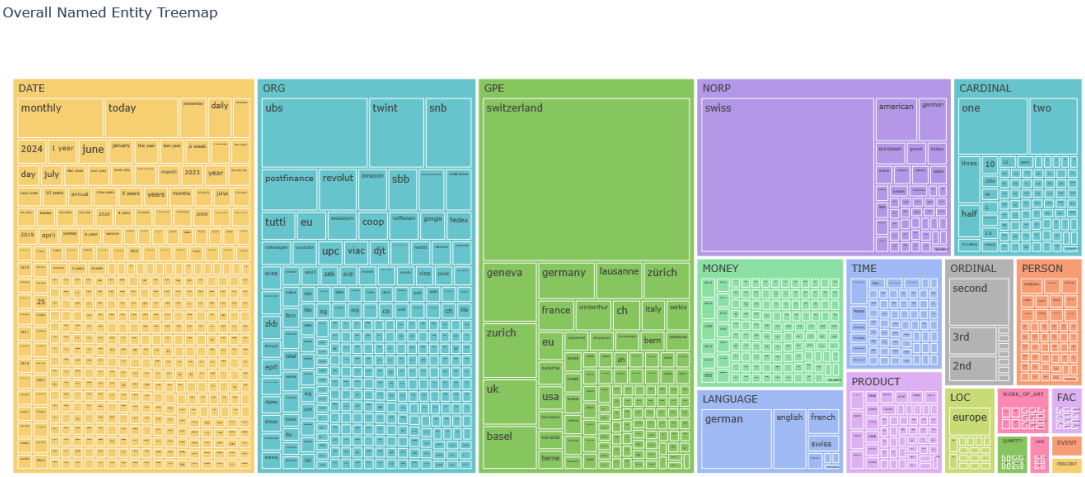

Figure 11: Hierarchical composition of named entity categories across the whole dataset (2010–2025). Primary categories are subdivided by their most frequent sub-terms, with area sizes proportional to occurrence counts. Colors denote entity categories. Source: Own illustration

The treemap (Figure 11) confirms DATE, ORG and GPE as the most dominant categories (consistent with Figure 10's frequency rankings), with their largest sub-terms (e.g., "monthly" under DATE and entities like "switzerland" under GPE) occupying the most visual space within their respective category. Lower-frequency categories (e.g., LANGUAGE, PRODUCT) appear as smaller tiles, mirroring their marginal representation in the corpus.

Lastly, Figure 12 presents the top 100 named entities as a word cloud, offering a more granular perspective on the dominant individual terms that define the dataset's composition. The visualization highlights term frequency through font scaling, with larger text indicating higher prevalence.

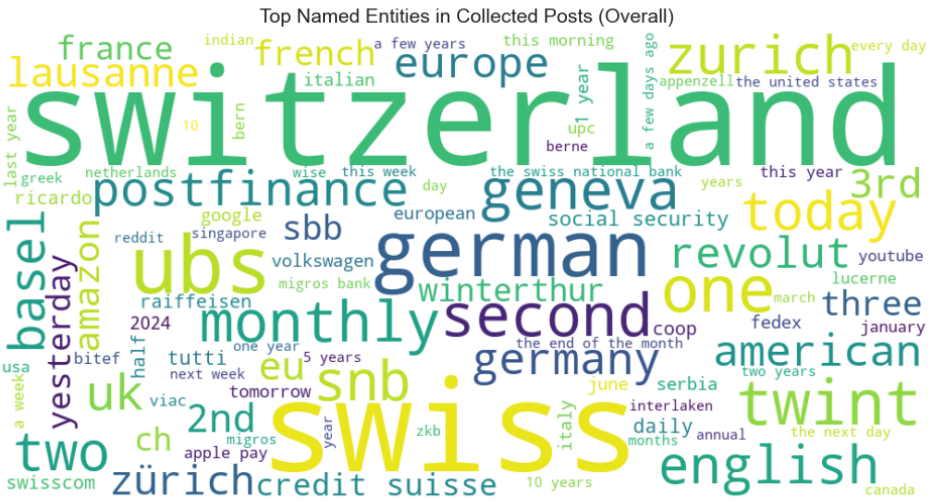

Figure 12: Word cloud visualization of the top 100 named entities extracted from collected posts across all years. Larger font sizes indicate higher frequency of mention. Prominent entities include geographic locations (e.g., "Switzerland", "Zürich"), financial institutions (e.g., "UBS", "PostFinance"), and common time expressions (e.g., "monthly", "today"). Source: Own illustration

The word cloud visualizations reinforce and further refine the insights obtained so far. While `DATE` entities dominate the dataset overall, they are less prominent in the word cloud due to their diversity: they appear frequently as a category but are dispersed across many distinct expressions. Instead, the most frequently mentioned entities are Switzerland (`GPE`) and Swiss (`NORP`), which, consistent with Figure 11, occupy the most visual space.

Also prominently featured are financial institutions such as *PostFinance*, *UBS*, *SNB*, and *Credit Suisse*, reflecting the high frequency of `ORG`-type entities noted earlier.

5.4 Topic Modeling Results

Results from topic modeling using BERTopic are limited. For individual years, the analysis was skipped either due to an insufficient number of posts or because all identified topics were classified as outliers. Only when aggregating posts across all years did BERTopic successfully generate interpretable results, identifying two distinct topics, which are presented below.

To provide an overview of how *Reddit* posts were distributed across the identified topics, Figure 13 presents a bar chart showing the number of documents assigned to each

topic. As BERTopic was only able to extract two valid topics from the combined dataset, the distribution reflects how the entire corpus was partitioned. The size of each topic indicates its relative prominence within the collected posts.

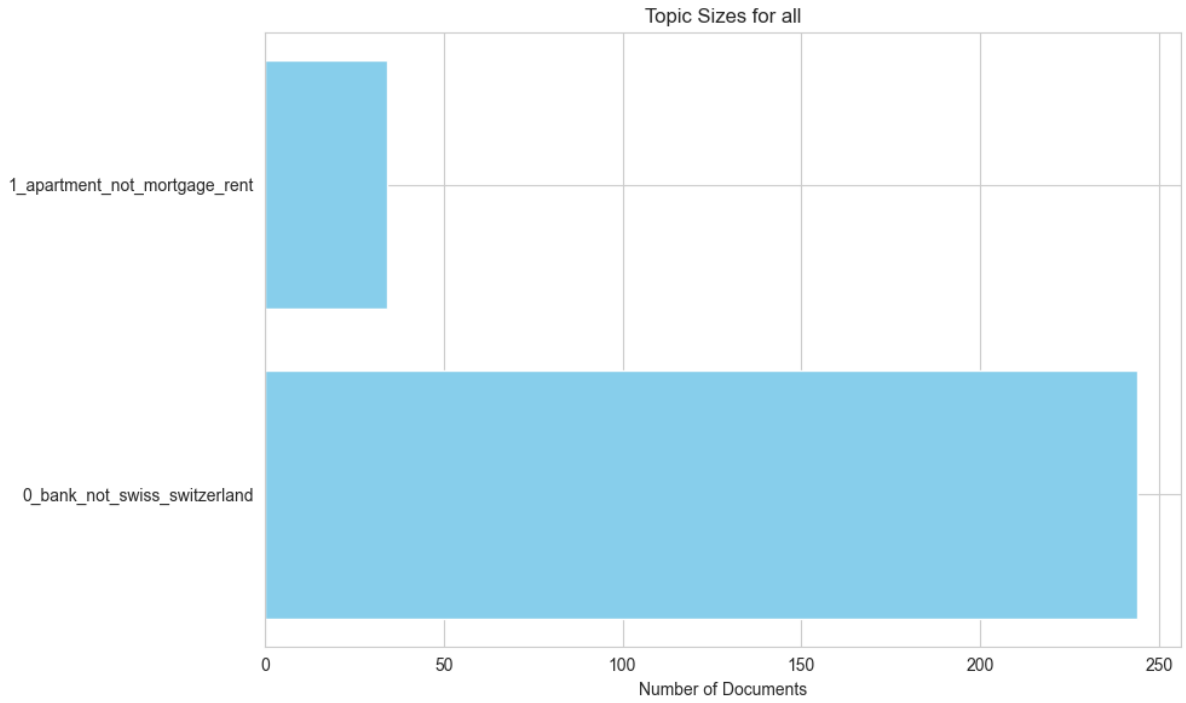

Figure 13: Bar chart showing the number of documents assigned by BERTopic to each extracted topic. Source: Own illustration

As shown, the distribution of documents across topics is highly imbalanced: the majority of posts (approximately 240) were assigned to topic 0, while a way smaller portion was classified under topic 1.

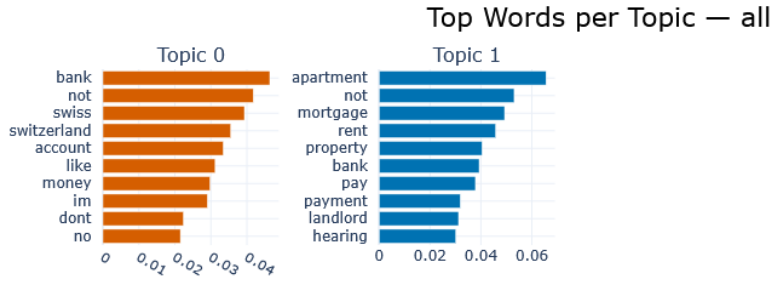

Figure 14: Bar charts generated by BERTopics `visualize_barchart()` method showing the top keywords associated with each identified topic (across all years combined). Source: Own illustration

To better understand the semantic content of each identified topic, Figure 14 presents

bar charts generated using BERTopic's built-in `visualize_barchart()` method. It displays the top keywords most strongly associated with each topic. These keywords offer insight into the main themes and concepts underlying each cluster of documents.

Topic 0 is characterized by financial terms such as "bank", "account", and "money", as well as negations like "not", "no", and "dont". Topic 1 includes keywords related to housing and rental themes, including "apartment", "mortgage", "property", and "rent". Similarly to topic 0, it also contains the negation "not". This presence of negations in both topics points towards potential discussions around financial or housing challenges and disagreements.

To illustrate the degree of similarity between these two discovered topics, Figure 15 shows a Venn diagram comparing their respective keyword sets. Shared keywords appear in the overlapping section, highlighting conceptual intersections, while distinct terms point to thematic differences between the two clusters. This view allows for interpretation of topic separation and cohesion.

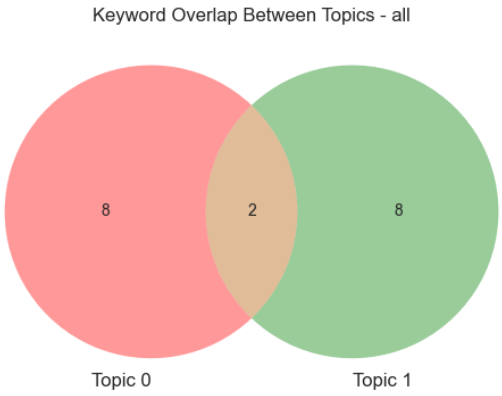

Figure 15: Venn diagram showing keyword overlap between the two discovered topics. Source: Own illustration

The Venn diagram shows that the topics share two keywords: "not" and "bank" (as seen in Figure 14). The remaining eight keywords in each topic are distinct, indicating that the topics are well separated in terms of their thematic content.

All results are available on GitHub, except for the BERTopic model, which is hosted on Google Drive due to file size constraints.

6 Discussion

This chapter discusses data quality considerations and reflects on the findings presented in the previous chapter, examining their implications in relation to the research questions, while identifying potential avenues for future research.

6.1 Data Quality

The dataset of 288 posts aligns with typical sample sizes in prior *Reddit* research as discussed in Section 2.1. While named entity recognition results confirm keyword effectiveness (e.g., frequent mentions of financial entities like "UBS" and "SNB" in Figure 12), keyword-based querying has inherent constraints: it may miss more nuanced economic discussions and some relevant content might have therefore been excluded (see Section 4.11.1). Since keywords alone cannot guarantee topical precision, the data may contain some noise. This is evidenced by BERTopic's inability to identify stable non-outlier topics on a yearly basis (except when data is aggregated), highlighting inherent noise limitations.

In conclusion, the data is consistent with keyword goals but suffers from methodological trade-offs. Pre-2020 data scarcity (<10 posts/year) renders sentiment trends highly unreliable and unstable (Section 5.2). Thus, analysis going forward focuses on the years from 2020 onward.

6.2 Year-by-Year Analysis

This section presents a year-by-year analysis of sentiment trends in *Reddit* posts and compares them to the key Swiss economic indicators introduced in Section 4.10. The year 2020 is established as a baseline, with each subsequent year analyzed in relation to the one before. The primary focus lies on measured sentiment (via VADER and FinBERT), while NER serves as a supporting lens to better understand the key actors and entities discussed in each year's posts.

Due to BERTopic's failure to generate meaningful topics for individual years, either because of insufficient data or because all topics were classified as outliers, its results are excluded from this section.

6.2.1 2020 - Baseline

In 2020, sentiment within *Reddit* posts related to the Swiss economy, as measured by VADER, was moderately positive at around **0.4**, while FinBERT returned a near-neutral score of approximately **-0.05** (Figure 8). As shown in Figure 16, the discourse was dominated by `GPE`s (geo-political entities), with Figure 17 indicating a focus on financial institutions (e.g., "UBS", "Raiffeisen"), events ("Art Basel"), car manufacturers ("Porsche", "BMW"), and countries such as "Switzerland", "The United States", and "South Africa".

Since 2020 serves as the baseline year, it provides the reference point for interpreting subsequent changes. The selected economic indicators (GDP, the SMI, CPI, and the Consumer Sentiment Index) are inherently comparative in nature and only gain analytical value when examined in terms of their year-over-year differences. As such, no evaluation of these indicators is provided for 2020 alone.

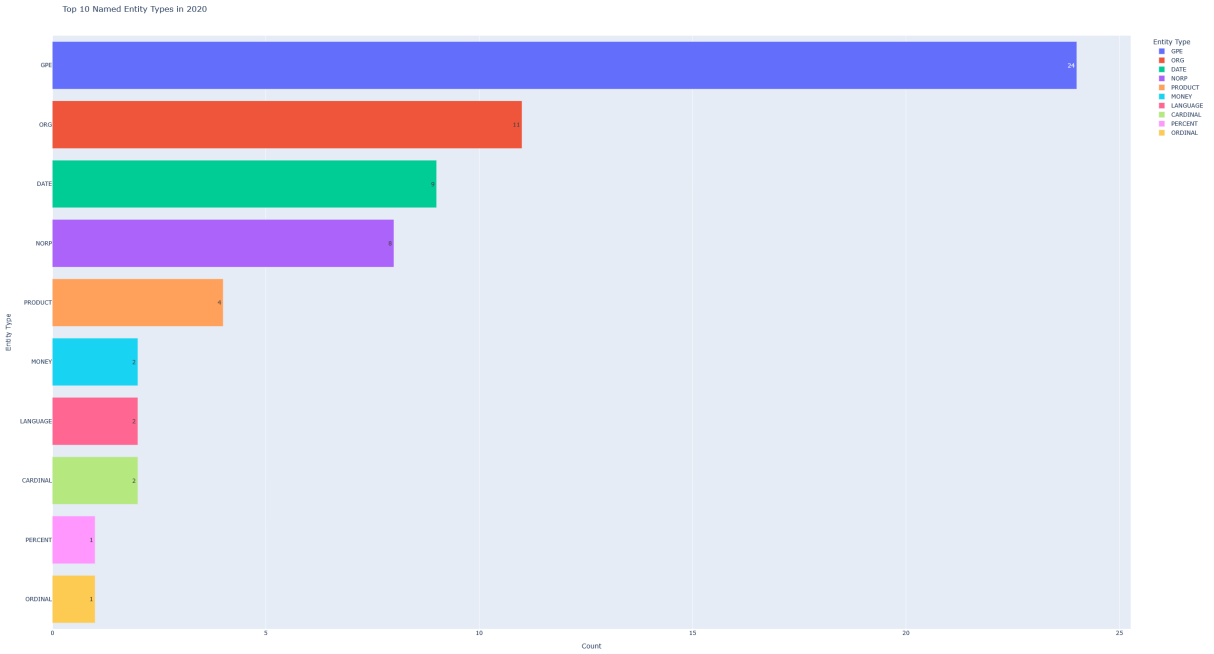

Figure 16: Frequency distribution of the top 10 named entity types for 2020. Values represent total counts per entity type. Source: Own illustration

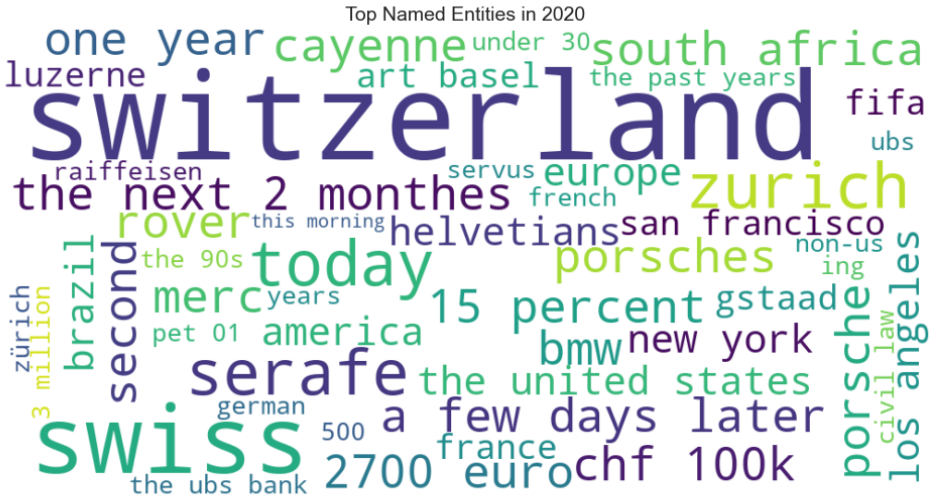

Figure 17: Word cloud visualization of top 100 named entities extracted from collected posts for 2020. Larger font sizes indicate higher frequency of mention. Source: Own illustration

6.2.2 2021

For 2021, sentiment as measured by VADER showed a moderate increase to approximately **0.6** (Figure 8), while FinBERT sentiment also rose slightly, though the change was minimal. Figure 18 indicates that the discourse was dominated by `DATE` entities, followed closely by `GPE`s and `NORP` (nationalities, religious, or political groups). Figure 19 shows that extracted entities once again included financial institutions such as "PostFinance" and "Revolut", alongside a notable presence of date references. This is consistent with the entity type distribution observed. Of further note is the presence of countries such as "UK" and "Thailand".

During this year, Switzerland's gross domestic product (GDP) increased by **7%** at current prices and **5.6%** in real terms compared to 2020. The annual inflation rate rose by **0.6%**, and the Swiss Market Index (SMI) ended the year at **12,875.66** points, up from **10,703.51** in 2020, an increase of approximately **20%**. Meanwhile, the consumer sentiment index averaged **-11.36**, up from **-27.5** in 2020.

Overall, these indicators point towards an economic recovery in Switzerland following the COVID-related downturn of 2020. The sharp GDP growth, a strong stock market performance, low but positive inflation, and a significantly improved consumer sentiment index all suggest renewed economic momentum. This context aligns well with the rise in

sentiment detected in *Reddit* discussions, particularly in the VADER results, indicating that online discourse has mirrored the broader economic optimism of the time.

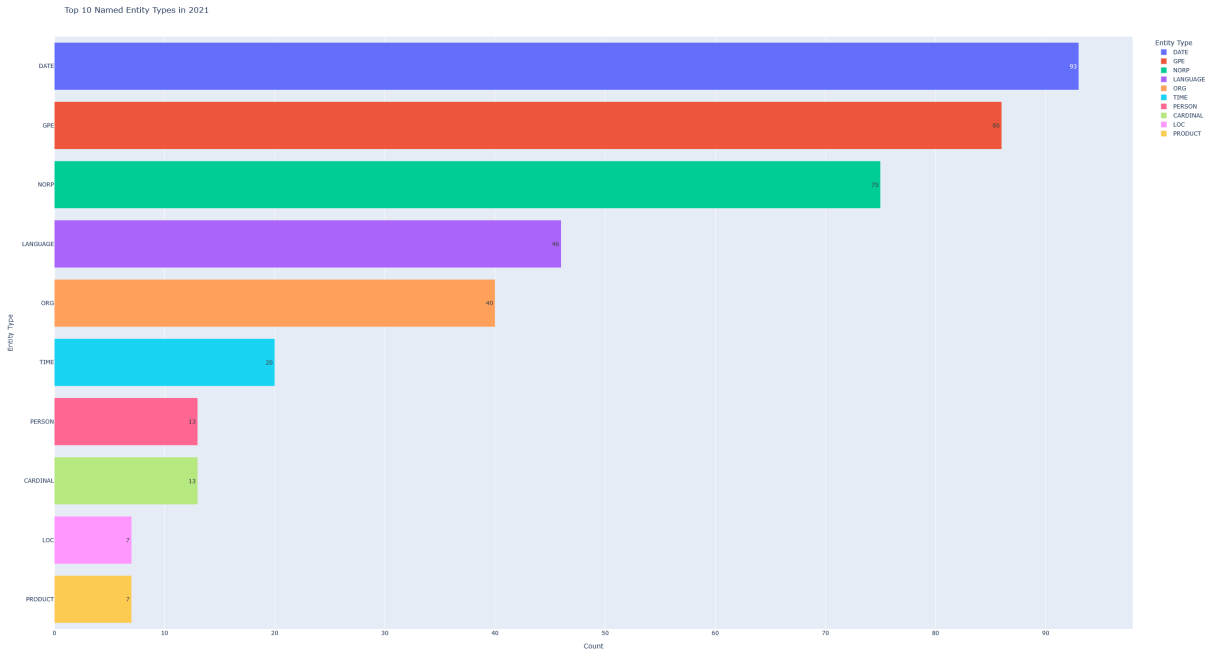

Figure 18: Frequency distribution of the top 10 named entity types for 2021. Values represent total counts per entity type. Source: Own illustration

Figure 19: Word cloud visualization of top 100 named entities extracted from collected posts for 2021. Larger font sizes indicate higher frequency of mention. Source: Own illustration

The prevalence of `DATE`, `GPE`, and `NORP` entities, alongside references to financial institutions such as *PostFinance* and *Revolut*, may suggest that users were discussing time-

specific developments or international and institutional aspects of the economy in relation to the post–COVID-19 recovery phase in 2021. The presence of geopolitical mentions like "UK" and "Thailand" further supports the interpretation that international contexts played a role in shaping these discussions.

6.2.3 2022

In 2022, as shown in Figure 8, VADER sentiment declined to slightly below **0.5**, while FinBERT sentiment also decreased marginally, returning to levels observed in 2020. According to Figure 20, *Reddit* discourse continued to be dominated by `DATE`-type entities, similar to the previous year, followed by `ORG` (organizations) and `GPE`s. Figure 21 supports this distribution, highlighting frequent mentions of dates, financial institutions, and companies such as "FedEx", "UPC" and "Aldi". Notably, the term "social security" also appears prominently, potentially reflecting public discourse around economic stability or welfare.

In terms of economic indicators, Switzerland's GDP increased by **6.2%** at current prices and **3%** at prices of the previous year, representing continued growth, though at a slower pace than in 2021. The annual inflation rate rose significantly by **2.8%**, while the SMI closed the year at **10,729.4** points, a decline that brought it nearly back to 2020 levels. The Consumer Sentiment Index averaged **−38.97** in 2022, representing a notable deterioration in public outlook.

While the GDP continued to grow in 2022, the sharp rise in inflation, decline in the stock market, and the lowest consumer sentiment value observed so far indicate growing concerns within the Swiss economy. This shift is reflected in the sentiment analysis: VADER shows a significant decline compared to 2021, while FinBERT sentiment falls slightly, returning to 2020 levels.

The frequent mentions of `DATE`, `ORG`, and `GPE` entities, alongside frequent mentions of financial institutions, corporations, and terms such as "social security", may indicate that discussions centered on institutional responses to economic pressures, rising living costs, or broader uncertainty, possibly influenced by the ripple effects of Russia's invasion of Ukraine.

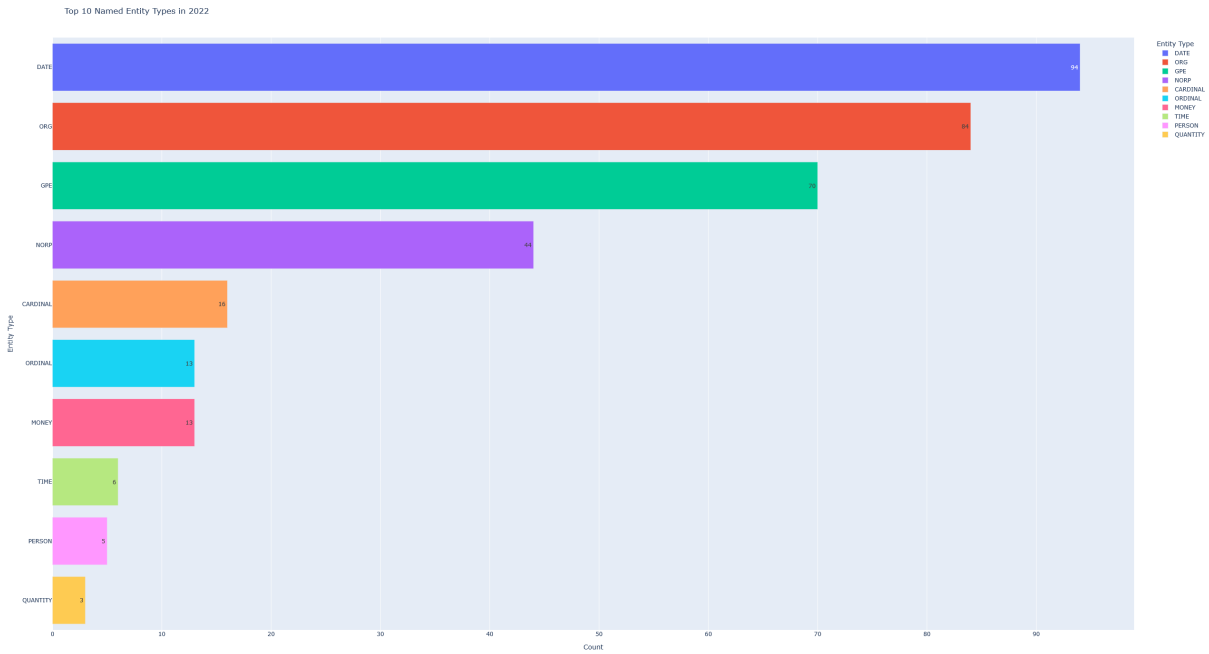

Figure 20: Frequency distribution of the top 10 named entity types for 2022. Values represent total counts per entity type. Source: Own illustration

Figure 21: Word cloud visualization of top 100 named entities extracted from collected posts for 2022. Larger font sizes indicate higher frequency of mention. Source: Own illustration

6.2.4 2023

VADER sentiment continued its downward trend in 2023, reaching approximately **0.35**, the lowest value observed so far. In contrast, FinBERT sentiment increased slightly,

recording its highest score to date, though still marginally negative overall (Figure 8). Notably, for the first time since 2020, `DATE` entities were no longer dominant. Instead, organizational (`ORG`) entities took precedence, followed by `DATE` and geopolitical (`GPE`) entities (Figure 22). This distribution is reflected in the word cloud (Figure 23), displaying financial organizations such as "Kantonalbank", "Twint", "PostFinance", and "Revolut". Also noteworthy is the appearance of geopolitical actors like "Russia" and "The US government".

When it comes to economic indicators, Switzerland's GDP grew by just **1.6%** at current prices and **0.7%** at prices of the previous year, indicating only marginal economic expansion. The average annual inflation rate increased by **2.1%**, slightly less than in 2022. The SMI stood at **11,137.79** points at the end of the year, a modest recovery from the previous year's decline. However, average consumer sentiment deteriorated further, reaching **−40.73**, an even lower value than in 2022.

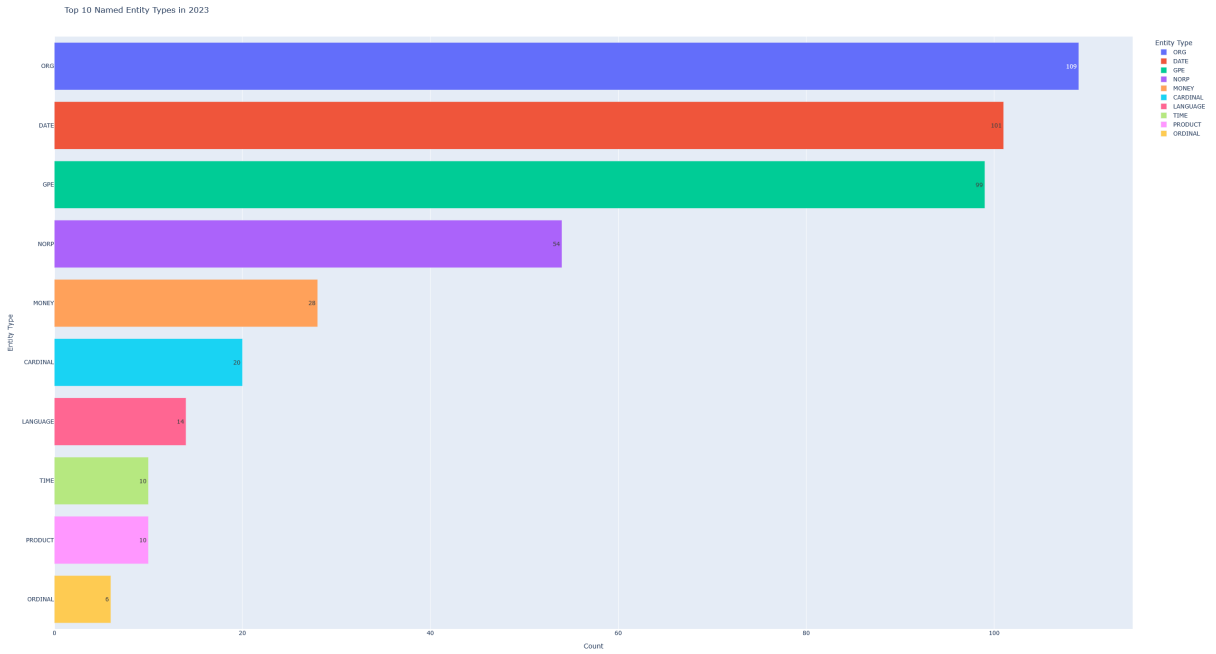

Figure 22: Frequency distribution of the top 10 named entity types for 2023. Values represent total counts per entity type. Source: Own illustration

The economic indicators for 2023 suggest a continued weakening of economic momentum in Switzerland. GDP growth was marginal, inflation remained elevated, and the consumer sentiment index reached its most negative value yet. Although the Swiss Market Index showed a modest rebound, this was not mirrored in public perception. Sentiment analysis supports this view: VADER sentiment declined further, reaching its lowest value

so far, while FinBERT sentiment increased slightly but remained negative overall.

The shift in named entity patterns, particularly the prominence of `ORG` entities over `DATE` ones, alongside frequent mentions of financial institutions such as *Kantonalbank*, *PostFinance*, and *Revolut*, as well as geopolitical references like "Russia" and "The US government", may suggest increased attention to institutional actors and international developments. This could indicate that users were discussing ongoing global tensions, inflationary pressures, or shifts in financial policy as economic uncertainty persisted throughout 2023.

Figure 23: Word cloud visualization of top 100 named entities extracted from collected posts for 2023. Larger font sizes indicate higher frequency of mention. Source: Own illustration

6.2.5 2024

As shown in Figure 8, neither VADER nor FinBERT registered any meaningful change in sentiment for 2024, with both models maintaining values similar to those observed in 2023. A very slight increase in VADER sentiment can be observed upon close inspection, though it remains negligible. This stability contrasts with the more dynamic shifts seen in previous years. In terms of named entity recognition, `DATE`-type entities returned to dominance in 2024, overtaking organizations once again (Figure 24). This pattern is further reflected in the word cloud (Figure 25), which shows numerous references to specific dates. Continuing trends from previous years include, once again, frequent mentions of financial institutions, alongside increased visibility of companies such as "Volkswagen",

"Google", and "YouTube".

Regarding economic indicators, Switzerland's GDP was estimated to have grown marginally by **0.35%** in 2024. As this figure is based on quarterly data rather than an official annual report (see Section 4.10), direct comparisons with previous years should be made with caution. Nevertheless, it suggests minimal economic expansion during the year. The average annual inflation rate increased by **1.1%**, marking a notable decline from the previous year. Meanwhile, the Swiss Market Index continued its upward trend, closing at **11,600.9** points. Consumer sentiment showed a modest recovery, with the index averaging **−37.14**, slightly higher than in 2022, though still deeply negative overall.

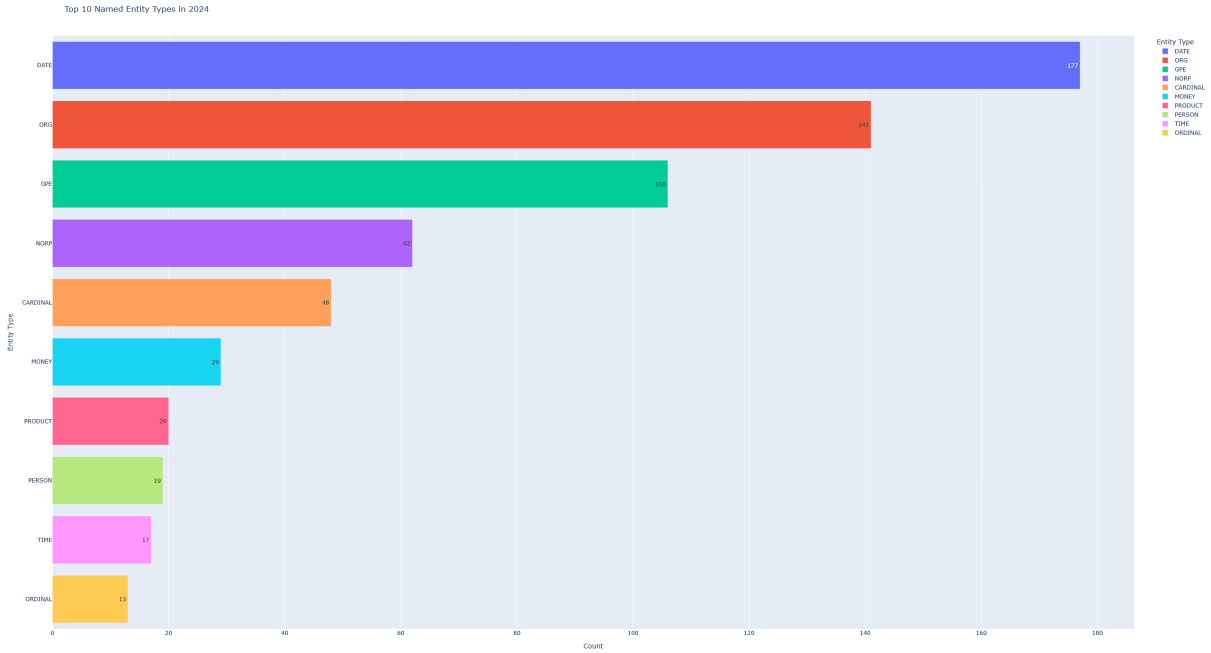

Figure 24: Frequency distribution of the top 10 named entity types for 2024. Values represent total counts per entity type. Source: Own illustration

This points to a tentative stabilization in Switzerland's economic environment. While GDP growth was minimal, inflation eased noticeably, and the Swiss Market Index continued its upward trend. Consumer sentiment remained low but showed slight improvement compared to the previous two years. This relative steadiness is echoed in the sentiment analysis: both VADER and FinBERT scores remained largely unchanged from 2023, with only a very slight increase in VADER sentiment. Rather than signaling a renewed sense of optimism, these results may reflect a pause in shifting public sentiment after several years of economic turbulence.

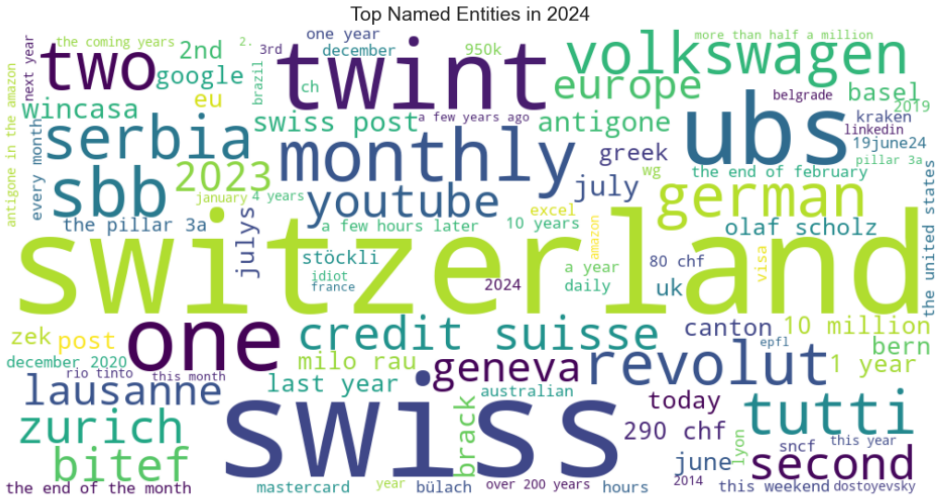

Figure 25: Word cloud visualization of top 100 named entities extracted from collected posts for 2024. Larger font sizes indicate higher frequency of mention. Source: Own illustration

In terms of named entity recognition, continued mentions of financial institutions and a variety of corporations such as *Volkswagen*, *Google*, and *YouTube*, combined with the dominance of `DATE` entities suggest that users remained focused on institutional actors and potentially time-sensitive developments. However, without significant shifts in sentiment, the discourse may have reflected a watchful or cautious stance rather than concern or optimism.

6.2.6 2025

In the ongoing year 2025, as shown in Figure 8, VADER and FinBERT sentiment scores moved in opposite directions: VADER sentiment increased, returning to levels last seen in 2020, while FinBERT sentiment declined sharply, reaching its lowest value. Organization-type entities once again became the most prevalent, overtaking `DATE` entities as the dominant category in *Reddit* discussions, similar to the pattern observed in 2023 (Figure 26). This dominance is further evidenced in Figure 27, with financial organizations appearing slightly more prominent than in the previous year. Notably, references to "Trump Media", "Trump", and the "EU" stand out, suggesting emerging geopolitical or political themes in the discourse.

As for the Swiss economic context, available data for 2025 remains incomplete and

should be interpreted cautiously. GDP grew by **0.8%** in the first quarter, while the average inflation rate so far this year has been estimated to increase by **0.2%**. The SMI closed at **11,836** points on July 31st, 2025, continuing its upward trend from the previous year. Meanwhile, the average consumer sentiment index was calculated to be **−37.31**, representing a very slight decline compared to 2024. However, as data is only partially available and may be revised later in the year, these figures offer only a preliminary view of the current economic climate.

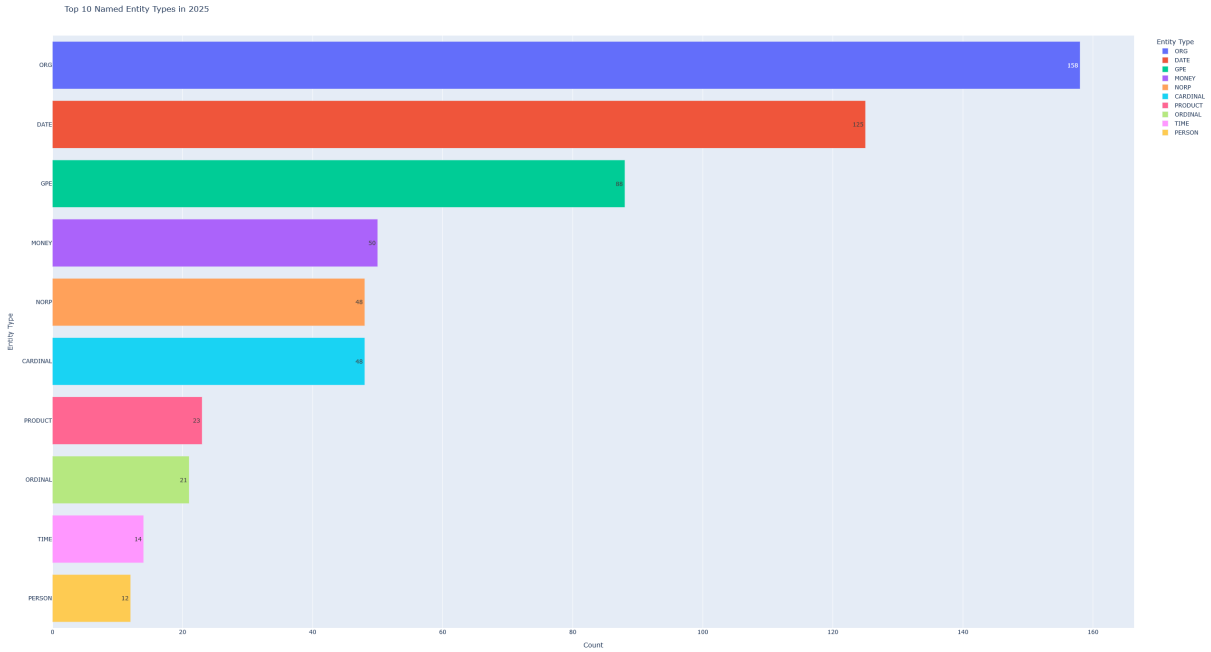

Figure 26: Frequency distribution of the top 10 named entity types for 2025. Values represent total counts per entity type. Source: Own illustration

These figures hint at a continuation of slow GDP growth and subdued inflation, while the Swiss Market Index has continued its upward trajectory. In contrast, consumer sentiment remained low, with a slight decline from 2024, suggesting that improvements in market indicators may not be fully mirrored in public perception. Sentiment analysis presents a mixed picture: VADER sentiment increased modestly, while FinBERT sentiment dropped. This divergence may reflect the fact that the year is still ongoing and could settle once full-year data becomes available.

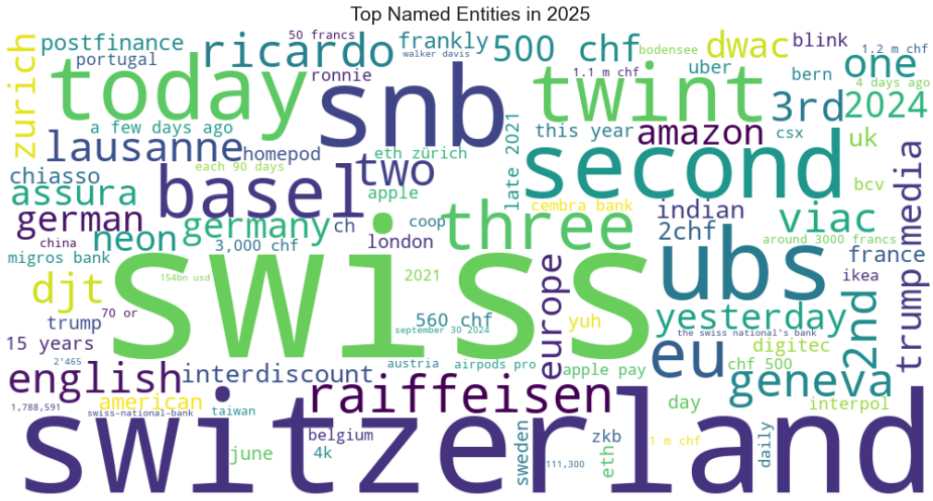

Figure 27: Word cloud visualization of top 100 named entities extracted from collected posts for 2025. Larger font sizes indicate higher frequency of mention. Source: Own illustration

The continued presence of financial institutions within NER results remained consistent, with slightly increased visibility compared to the previous year. Noteworthy are mentions of "Trump", "Trump Media", and the "EU", which may indicate growing attention to international political developments or media-driven discourse. Given the provisional nature of the data and lack of clear sentiment trends, 2025 may represent a transitional period, marked more by uncertainty than by strong economic or emotional shifts in public discourse.

6.3 Conclusion

The year-by-year analysis presented in this chapter examined how sentiment expressed in *Reddit* posts compared to Swiss economic developments from 2020 to 2025. Sentiment analysis via VADER and FinBERT served as the primary analytical lens, while named entity recognition contributed supporting context by highlighting the key entities and institutions driving public discourse. Together, these methods helped surface patterns between online discussions and official economic indicators.

One additional methodological subtlety deserves attention, particularly in relation to the data quality considerations raised earlier in Section 6.1. Some yearly word clouds (e.g., Figure 19; Figure 23) contain terms that are not in English, such as "Hochzeit-

sregister" and "Lohnkonten". This suggests that the language filtering implemented in the data collection script (Section 4.3) may not have performed optimally in all cases. Alternatively, these terms may have appeared within otherwise English-language posts, used deliberately due to their specificity or lack of direct translation. In either case, this highlights the presence of linguistic heterogeneity in Swiss online discourse and points to potential refinements in language detection for future studies.

Overall, the analysis confirms that sentiment expressed by users in economy-focused *Reddit* posts from r/Switzerland largely mirrors real-world economic developments as captured by key national indicators, thereby **supporting the first part of the research question**. This alignment was particularly evident in the years 2021 through 2024. For 2025, however, no definitive conclusion can yet be drawn, as full-year data remains unavailable and the sentiment trends between the two applied models diverge.

This pattern was especially pronounced in the VADER results, which consistently tracked with shifts in GDP growth, inflation, stock market trends, and consumer sentiment. FinBERT, by contrast, exhibited more minimal variation and occasional divergence, particularly in 2023 and 2025. These findings suggest that while FinBERT may be well-suited to formal financial texts, it appears less responsive to the informal and contextually diverse nature of *Reddit* discourse, even when that discourse is financially themed. This may indicate limitations in FinBERT's previously discussed robustness to slight shifts in domain and a lack of ability to adapt to colloquial, user-generated, short-form content.

Regarding named entity recognition, one notable trend was the recurring appearance of financial institutions, such as banks and payment providers, across all examined years. This pattern is likely influenced by the initial keyword selection strategy used during data collection. Given that the query terms used explicitly targeted economic and financial topics and even referenced such entities directly (e.g., "SNB", "UBS", and "Credit Suisse"), the frequent mentions of these institutions are unsurprising. While their sustained presence aligns with the thematic focus of this study, it should not be interpreted as emergent from user discourse. Rather, their prominence partially reflects the structure of the dataset itself. That said, their continued visibility, especially when observed alongside more context-specific or temporally dynamic entities, as discussed throughout the yearly analyses, does suggest that institutional actors remained consistently central in public

discussions of economic matters over the examined time period. Beyond their frequency, the named entity recognition results also helped contextualize sentiment trends by highlighting the specific terms around which discussions were centered, thus offering insights into the substance of public discourse, rather than sentiment alone. In this way, NER helped establish what was being discussed, complementing sentiment analysis by adding thematic and narrative texture to its findings.

BERTopic, by contrast, offered only limited analytical value. It failed to extract stable or interpretable topics on a year-by-year basis, largely due to insufficient post volume or overly heterogeneous and noisy content. While two general topics could be identified for the entire dataset, they were too coarse to support fine-grained temporal analysis. As such, topic modeling was excluded from this chapter's core findings and may require alternative strategies or more specific data in future work.

As for the second part of the research question - whether *Reddit* sentiment could serve as a predictive signal for economic developments - **no evidence to support that claim could be found**. During the observed period (2020–2025), sentiment trends did not consistently lead or lag behind official economic indicators. It is possible that a higher temporal resolution (monthly or quarterly rather than yearly) could help uncover short-term predictive patterns. However, given the limited number of posts per year (peaking at just 67 in 2024), the data volume at such levels would likely be insufficient for robust analysis under the current methodology.

7 Conclusion & Future Work

This thesis developed an NLP pipeline to analyze economy-focused *Reddit* posts from the subreddit r/Switzerland. The core objective was to assess whether user-expressed sentiment reflected Swiss economic developments, and whether such sentiment could serve as a predictive signal. The analysis combined sentiment analysis (VADER, FinBERT) with transformer-based named entity recognition (SpaCy) and topic modeling (BERTopic) to contextualize discourse.

Analysis revealed that sufficient data was only available from 2020 onward. Consequently, 2020 was established as the baseline year, with each following year analyzed in comparison to the previous one.

Sentiment analysis showed that sentiment values, particularly those generated by VADER, closely tracked key economic indicators, including gross domestic product (GDP), the consumer price index (CPI), the Swiss Market Index (SMI), and the Consumer Sentiment Index. In contrast, FinBERT was found to be less effective when applied to the informal and colloquial nature of *Reddit* content, exhibiting minimal variation and occasional divergence. Throughout the study period, sentiment did not exhibit predictive power with respect to economic trends.

Named entity recognition served a supporting role by highlighting key terms and actors within each year's discussions, providing additional context to the sentiment trends observed. In contrast, results from BERTopic were limited: the model was unable to extract topics on a yearly basis and only identified coherent topics when all years were aggregated.

The results of this thesis are broadly consistent with prior research (Section 2.2) that has demonstrated correlations between social media sentiment and real-world economic developments across various platforms. Similar to the studies conducted by Fano and Toschi (2022) on *Twitter* data and Pan and Li (2015) on *Weibo*, VADER-based sentiment analysis of r/Switzerland posts aligned with key economic indicators, particularly between 2021 and 2024. However, in contrast to some earlier work reporting predictive utility of online sentiment (e.g., Durai and Wang, 2023; Chen and Ma, 2024), this study found no evidence that *Reddit* sentiment led or lagged official indicators. This difference may reflect the yearly temporal resolution, the smaller dataset size, and the more generalist (rather than finance-specific) nature of the studied subreddit.

Notable limitations of this study include the behavior of the *Reddit* API, which was found to return inconsistent results, even for identical queries submitted seconds apart. Additionally, the keyword list used may have introduced trade-offs: Expanding and refining the list past a certain point reduced the number of API results, while still not ensuring full topical accuracy in the collected posts. As a result, some noise in the dataset is likely and more nuanced economy-focused content might have been excluded. This noise is reflected, for example, in BERTopic's inability to extract meaningful yearly topics. Finally, close inspection of the results revealed occasional inconsistencies in language filtering, suggesting that the implemented detection method did not always perform optimally. Lastly, the temporal resolution of the analysis was limited, as sentiment was evaluated on

a yearly basis rather than at a finer granularity, such as monthly. As a result, the findings of this study should be interpreted with these limitations in mind.

Addressing these constraints presents several avenues for future work, particularly through efforts to denoise the data. One approach could involve usage of an LLM to verify the thematic accuracy of collected posts, instead of relying solely on keyword-based querying. Additionally, more robust language filtering techniques could be implemented to minimize German words interfering with NER results. Improving temporal resolution, potentially to monthly or even daily intervals, might enable the extraction of predictive (rather than merely correlative) patterns, though this would require sufficient data, possibly obtained by expanding the scope to include multiple subreddits (e.g., r/Europe, r/Economics). Finally, model performance could be enhanced, for instance, by fine-tuning FinBERT on colloquial *Reddit* content to improve sentiment analysis accuracy. Collectively, these refinements could strengthen the link between online sentiment and real-world economic developments, and better assess its potential as a predictive signal.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131. https://doi.org/https://doi.org/10.1016/j.is.2022.102131

Aihkisalo, T., & Paaso, T. (2011). A performance comparison of web service object marshalling and unmarshalling solutions. *2011 IEEE World Congress on Services*, 122–129. https://doi.org/10.1109/SERVICES.2011.61

Alrashed, T., Almahmoud, J., Zhang, A. X., & Karger, D. R. (2020). Scrapir: Making web data apis accessible to end users. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376691

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. https://arxiv.org/abs/1908.10063

Araci, D., & Genc, Z. (2022). *ProsusAI/finBERT: Financial Sentiment Analysis with BERT*. Retrieved June 28, 2025, from https://github.com/ProsusAI/finBERT

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in twitter sentiment analysis. In D.-S. Huang, K.-H. Jo, & L. Wang (Eds.), *Intelligent computing methodologies* (pp. 615–624). Springer International Publishing.

Bauer, T., Beer, F., Holl, D., Imeraj, A., Schweiger, K., Stangl, P., Weigl, W., & Neumann, C. (2022, July). *Reddiment: Eine sveltekit- und elasticsearch- basierte reddit sentiment-analyse* (No. CL-2022-06). Ostbayerische Technische Hochschule Amberg-Weiden, CyberLytics-Lab an der Fakultät Elektrotechnik, Medien und Informatik. https://doi.org/10.13140/RG.2.2.32244.12161

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. https://arxiv.org/abs/2001.08435

Bieri, M.-C. (2023). Assessing economic sentiment with newspaper text indices: Evidence from switzerland. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4672018

Boe, B. (2023a). *Praw 7.7.1 documentation*. Retrieved June 7, 2025, from https://praw.readthedocs.io/en/stable/

Boe, B. (2023b). *Praw 7.7.1 documentation - listinggenerator*. Retrieved June 7, 2025, from https://praw.readthedocs.io/en/stable/code_overview/other/listinggenerator.html#praw.models.ListingGenerator

Callen, T. (2008). What is gross domestic product. *Finance & Development*, *45*(4), 48–49.

Carlucci, R. (2024). *Structure and dynamics of discussions on the social media platform reddit* [Ph.D. dissertation]. Georg-August-Universität Göttingen. https://doi.org/10.53846/goediss-10502

cash.ch. (2025). *Aktien smi*. Retrieved August 7, 2025, from https://www.cash.ch/aktien/smi-index-swl

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

Chen, Y., & Ma, H. (2024). *Predicting Stock Market Using Reddit Comments* [working paper or preprint]. https://hal.science/hal-04667405

Durai, S. A., & Wang, Z. (2023). Sentiment analysis, social media and urban economics: The case of singaporean hdb and covid-19. *International Journal of Innovation and Economic Development*, *9*(5), 28–39.

Eriksson, M., & Hallberg, V. (2011). *Comparison between json and yaml for data serialization* [Bachelor's Thesis]. The School of Computer Science and Engineering, Royal Institute of Technology.

Explosion AI. (2025a). *Entitiyrecognizer - spacy api documentation*. Retrieved June 28, 2025, from https://spacy.io/api/entityrecognizer

Explosion AI. (2025b). *Facts & figures - spacy usage documentation*. Retrieved June 15, 2025, from https://spacy.io/usage/facts-figures#benchmarks

Fano, S., & Toschi, G. (2022). Covid-19 pandemic and the economy: Sentiment analysis on twitter data. *International Journal of Computational Economics and Econometrics*, *1*, 1. https://doi.org/10.1504/IJCEE.2022.10046488

Federal Statistical Office (FSO). (2024). *Gross domestic product*. Retrieved July 31, 2025, from https://www.bfs.admin.ch/bfs/en/home/statistics/national-economy/national-accounts/gross-domestic-product.html

Federal Statistical Office (FSO). (2025). *Consumer prices*. Retrieved July 4, 2025, from https://www.bfs.admin.ch/bfs/en/home/statistics/prices/consumer-price-index.html

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* [Doctoral dissertation, University of California, Irvine].

Gilbert, S. A. (2020). "i run the world's largest historical outreach project and it's on a cesspool of a website." moderating a public scholarship site on reddit: A case study of r/askhistorians. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW1). https://doi.org/10.1145/3392822

Gjurković, M., Karan, V. M., Vukojević, I., Bošnjak, M., & Snajder, J. (2021, June). PANDORA talks: Personality and demographics on Reddit. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the ninth international workshop on natural language processing for social media* (pp. 138–152). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.socialnlp-1.12

Grochowski, K., Breiter, M., & Nowak, R. (2019). Serialization in object-oriented programming languages. In K. Sud, P. Erdogmus, & S. Kadry (Eds.), *Introduction to data science and machine learning*. IntechOpen. https://doi.org/10.5772/intechopen.86917

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. https://arxiv.org/abs/2203.05794

Grootendorst, M. (2024). *Bertopic*. Retrieved June 28, 2025, from https://maartengr.github.io/BERTopic/index.html

Hericko, M., Juric, M. B., Rozman, I., Beloglavec, S., & Zivkovic, A. (2003). Object serialization analysis and comparison in java and .net. *SIGPLAN Not.*, *38*(8), 44–54. https://doi.org/10.1145/944579.944589

Hunt, J. (2023). *A beginners guide to python 3 programming* (2nd ed.). Springer Cham. https://doi.org/10.1007/978-3-031-35122-8

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Jatana, N., Puri, S., Ahuja, M., Kathuria, I., & Gosain, D. (2012). A survey and comparison of relational and non-relational database. *International Journal of Engineering Research & Technology*, *1*(6), 1–5.

Jungherr, A., Posegga, O., & An, J. (2022). Populist supporters on reddit: A comparison of content and behavioral patterns within publics of supporters of donald trump

and hillary clinton. *Social Science Computer Review*, *40*(3), 809–830. https://doi.org/10.1177/0894439321996130

K, G., Chintalapati, A., Senapati, A., & Enkhbat, K. (2024). Sentiment analysis on reddit trading data. *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, 1–7. https://doi.org/10.1109/icETITE58242.2024.10493402

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68. https://doi.org/https://doi.org/10.1016/j.bushor.2009.09.003

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web scraping techniques and applications: A literature review. https://doi.org/10.52458/978-93-91842-08-6-38

Macale, S. (2011). *A rundown of reddit's history and community [infographic]*. Retrieved June 2, 2025, from https://thenextweb.com/news/a-rundown-of-reddits-history-and-community-infographic

Machavarapu, A. (2022). Reddit sentiments effects on stock market prices. In V. Bhateja, S. C. Satapathy, C. M. Travieso-Gonzalez, & T. Adilakshmi (Eds.), *Smart intelligent computing and applications, volume 1* (pp. 75–84). Springer Nature Singapore.

Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, *8*(2), 339–344.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. https://doi.org/https://doi.org/10.1016/j.csi.2012.09.004

Masse, M. (2011). *Rest api design rulebook: Designing consistent restful web service interfaces*. O'Reilly Media, Inc.

Mudassir, M., & Mushtaq, M. (2024). The role of apis in modern software development. *World Journal of Advanced Engineering Technology and Sciences*, *13*(1), 1045–1047. https://doi.org/10.30574/wjaets.2024.13.1.0515

NCRI / Pushshift. (2025). *Ncri reddit access*. Retrieved June 3, 2025, from https://pushshift.io/signup

Nikhila Kanigiri, S., Mekuriyaw, C., Goodman, G., & Alexiou, M. S. (2024). Analyzing the impact of preprocessing techniques on the efficiency and accuracy of sentiment classification algorithms. *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–8. https://doi.org/10.1109/IISA62523.2024.10786699

Nurseitov, N., Paulson, M., Reynolds, R., & Izurieta, C. (2009). Comparison of json and xml data interchange formats: A case study. *ISCA International Conference on Computer Applications in Industry and Engineering*. https://api.semanticscholar.org/CorpusID:16978698

Pan, F., & Li, H. (2015). Sina weibo mood predicting the economics trends. *Proceedings of the 2015 International Conference on Social Science, Education Management and Sports Education*, 428–431. https://doi.org/10.2991/ssemse-15.2015.108

Payne, J. (2024a). *Async praw 7.8.1 documentation*. Retrieved June 7, 2025, from https://asyncpraw.readthedocs.io/en/stable/

Payne, J. (2024b). *Async praw 7.8.1 documentation - listinggenerator*. Retrieved June 7, 2025, from https://asyncpraw.readthedocs.io/en/stable/code_overview/other/listinggenerator.html#asyncpraw.models.ListingGenerator

Pradha, S., Halgamuge, M. N., & Tran Quoc Vinh, N. (2019). Effective text data preprocessing technique for sentiment analysis in social media data. *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 1–8. https://doi.org/10.1109/KSE.2019.8919368

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, *7*(2), 20563051211019004. https://doi.org/10.1177/20563051211019004

Python Software Foundation. (2025a). *Json — json encoder and decoder*. Retrieved June 11, 2025, from https://docs.python.org/3/library/json.html

Python Software Foundation. (2025b). *Sqlite3 — db-api 2.0 interface for sqlite databases*. Retrieved June 11, 2025, from https://docs.python.org/3/library/sqlite3.html

Python Software Foundation. (2025c). *Vadersentiment - pypi*. Retrieved June 28, 2025, from https://pypi.org/project/vaderSentiment/

Quiña-Mera, A., Fernandez, P., García, J. M., & Ruiz-Cortés, A. (2023). Graphql: A systematic mapping study. *ACM Comput. Surv.*, *55*(10). https://doi.org/10.1145/3561818

Reddit Inc. (2023). *Data api terms*. Retrieved June 7, 2025, from https://redditinc.com/policies/data-api-terms

Reddit Inc. (2024). *Reddit inc homepage*. Retrieved June 2, 2025, from https://redditinc.com/

Reddit Inc. (2025a). *Brand foundation: Logo*. Retrieved June 2, 2025, from https://redditbrand.lingoapp.com/s/Logo-d9x3n2

Reddit Inc. (2025b). *Reddit api documentation*. Retrieved June 7, 2025, from https://www.reddit.com/dev/api/

Reddit Inc. (2025c). *Reddit api documentation - get [/r/subreddit]/search*. Retrieved June 7, 2025, from https://www.reddit.com/dev/api/#GET_search

Reddit Inc. (2025d). *Reddit data api wiki*. Retrieved June 7, 2025, from https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

Scherrmann, M. (2023). German finbert: A german pre-trained language model. https://arxiv.org/abs/2311.08793

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 338–343. https://doi.org/10.1109/SNAMS.2019.8931850

Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). Sentiment analysis techniques for social media data: A review. In A. K. Luhach, J. A. Kosa, R. C. Poonia, X.-Z. Gao, & D. Singh (Eds.), *First international conference on sustainable technologies for computational intelligence* (pp. 75–90). Springer Singapore.

Shelar, H., Kaur, G., Heda, N., & and, P. A. (2020). Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries*, *39*(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479

SIX Group. (2025). *Swiss market index (smi) - der index für den blue-chip-aktienmarkt der schweiz*. Retrieved July 4, 2025, from https://www.six-group.com/de/market-data/indices/switzerland/equity/smi.html?utm_campaign=vanity-url&utm_medium=redirect&utm_source=www.six-group.com/smi

Soni, A., & Ranga, V. (2019). Api features individualizing of web services: Rest and soap. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9), 664–671.

SQLite. (2025a). *Atomic commit in sqlite*. Retrieved June 11, 2025, from https://www.sqlite.org/atomiccommit.html

SQLite. (2025b). *Most widely deployed sql database engine*. Retrieved June 11, 2025, from https://www.sqlite.org/mostdeployed.html

State Secretariat for Economic Affairs (SECO). (2025a). *Consumer sentiment*. Retrieved July 4, 2025, from https://www.seco.admin.ch/seco/en/home/wirtschaftslage---wirtschaftspolitik/Wirtschaftslage/Konsumentenstimmung.html

State Secretariat for Economic Affairs (SECO). (2025b). *Consumer sentiment - data from 1972 onwards*. Retrieved July 31, 2025, from https://www.seco.admin.ch/seco/en/home/wirtschaftslage---wirtschaftspolitik/Wirtschaftslage/Konsumentenstimmung/daten-ab-1972.html

State Secretariat for Economic Affairs (SECO). (2025c). *Gross domestic product*. Retrieved July 4, 2025, from https://www.seco.admin.ch/seco/en/home/wirtschaftslage---wirtschaftspolitik/Wirtschaftslage/bip-quartalsschaetzungen-.html

Statista. (2025). *Jährliche kursdaten des schweizer aktienindex swiss market index (smi) von 2000 bis 2024*. Retrieved July 31, 2025, from https://de.statista.com/statistik/daten/studie/972185/umfrage/jaehrliche-kursdaten-des-schweizer-aktienindex-smi/

Taylor, A. G. (2003). *Sql for dummies*. John Wiley & Sons.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/https://doi.org/10.1016/j.is.2020.101582

Vickery, G., & Wunsch-Vincent, S. (2007). *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD).

Wachirapong, F. (2023). *Exploring the effect of preprocessing techniques on the topic modeling in social science data* [Master's thesis, University of Helsinki]. http://hdl.handle.net/10138/572335

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Wankmüller, S. (2024). Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research*, *53*(4), 1676–1752. https://doi.org/10.1177/00491241221134527

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Weld, G., Glenski, M., & Althoff, T. (2021). Political bias and factualness in news sharing across more than 100,000 online communities. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 796–807. https://doi.org/10.1609/icwsm.v15i1.18104