

Beyond Twitter: Exploring Alternative API Sources for Social Media Analytics

Alina Campan^a and Noah Holtke

School of Computing and Analytics, Northern Kentucky University, Nunn Drive, Highland Heights, U.S.A.

Keywords: Social Media Analytics, Federated Social Media Platforms, API-Scraping.


Abstract: Social media is a valuable source of data for applications in a multitude of fields: agriculture, banking, business intelligence, communication, disaster management, education, government, health, hospitality and tourism, journalism, management, marketing, etc. There are two main ways to collect social media data: web scraping (requires more complex custom programs, faces legal and ethical concerns) and API-scraping using services provided by the social media platform itself (clear protocols, clean data, follows platform established rules). However, API-based access to social media platforms has significantly changed in the last few years, with the mainstream platforms placing more restrictions and pricing researchers out. At the same time, new, federated social media platforms have emerged, many of which have a growing user base and could be valuable data sources for research. In this paper, we describe an experimental framework to API-scrape data from the federated Mastodon platform (specifically its flagship node, Mastodon.social), and the results of volume, sentiment, emotion, and topic analysis on two datasets we collected – as a proof of concept for the usefulness of sourcing data from the Mastodon platform.

1 INTRODUCTION

Social media and online social networks (OSNs) have been a primary means to spread and consume information for a while now, due to the low cost and high pervasiveness. Despite their negative aspects, such as the echo chamber effect, and their potential for the spread of misinformation and disinformation, the discourse on social media also has positive dimensions, as is reflective of real-world events and trends. This allows for the positive use of social media data in a multitude of application fields: agriculture, banking, business intelligence, communication, disaster management, disruptive technology, education, ethics, government, health, hospitality and tourism, journalism, management, marketing, understanding terrorism (Zachlod, 2022). Different analysis methods are being used, including sentiment analysis, topic discovery, word frequency analysis and content analysis (Zachlod, 2022); also, analysis methods are still being researched and developed that are capable of effectively handling massive amounts of social media data (Zachlod,

2022) with acceptable accuracy. Commercial tools for social media analysis are also available.

Despite the variety of application fields and analytic methods, the deployment of social media analysis frameworks follows similar “steps necessary to gain useful information or even knowledge out of social media”; these steps are discovery, tracking (or collection), preparation, and analysis (Stieglitz, 2018), (Zachlod, 2022). In the tracking step, data is collected from one (or more) social media platform(s), using the provided communication method (API, RSS, HTML scraping.) In a recent literature review, Zachlod reported that from 94 articles they reviewed, the social media platforms investigated in these research works were: Twitter (55 studies), Facebook (25 studies), Instagram (13 studies), YouTube (8 studies), TripAdvisor (8 studies), LinkedIn (4 studies), other - Foursquare, Google +, TikTok, WeChat, Sina Weibo (21 times) (Zachlod, 2022). Of all social media sites, Twitter used to be the most popular. Twitter was once the dominant social media platform among all others. This was due to its less stringent privacy controls compared to platforms like Facebook (as Twitter is a

^a <https://orcid.org/0000-0002-9296-3036>

microblogging site designed for widespread dissemination of opinions, rather than communicating within a small group of friends), and in no small measure to its free API access for researchers. However, in recent years, there has been a shift towards a monetized model. Now, researchers must pay \$100 for a subscription that permits sampling of only 10,000 messages per month. Given the uncertain future of Twitter's accessibility for academic research, an investigation of alternative social media sources and their potential for research is worth investigating. In this work, we are considering one of the new social media platforms, the federated Mastodon platform, and specifically its flagship node, Mastodon.social. We explored its API technology, scraped two datasets focused on two different topics for a short window of time, analyzed the daily volume, sentiment valence, and emotional content of the two datasets – as a proof of concept for the usefulness of sourcing data from the Mastodon platform.

2 RELATED WORK: MASTODON.SOCIAL

Mastodon is a social media service that has become popular as an alternative to X (formerly Twitter) since its inception in 2016. Users engage with the platform by posting short-form content and engaging in conversations in comment feeds. Communities consist of self-hosted servers, often owned and operated by users of the platform, which integrate fully with all other Mastodon servers in the network. Each of these "instances" maintains its own community standards, policies, and content moderation. Users are free to join whichever instance they choose, but their account retains the ability to browse all public content on the platform. This federated model of content hosting has contributed to the development of a diverse range of communities.

The acquisition of Twitter by Elon Musk has contributed to a significant increase in the size of Mastodon's user base, growing by as many as 700,000 accounts between October and December in 2022. Its active user base currently sits around 1 million active users, with historic highs around 2.5 million. Accounting for a constant influx of new publications, a high volume of decentralized instances, and the distinction between public and private content, it is difficult to estimate the total volume of unique content on Mastodon.

“Mastodon.social is one of the many independent Mastodon servers you can use to participate in the

fediverse” (Mastodon.social, 2024). Mastodon.social is considered the flagship instance and sits currently (July 13, 2024) at 226K active users.

The figure below shows the evolution of numbers of Mastodon servers, users, and active; these statistics are available online at (Khun, 2024). The most recent statistics from (Mastodon statistics, 2024) show the numbers of servers at 9168, users at 8,741,802, and active users at 854,905, on July 13, 2024.

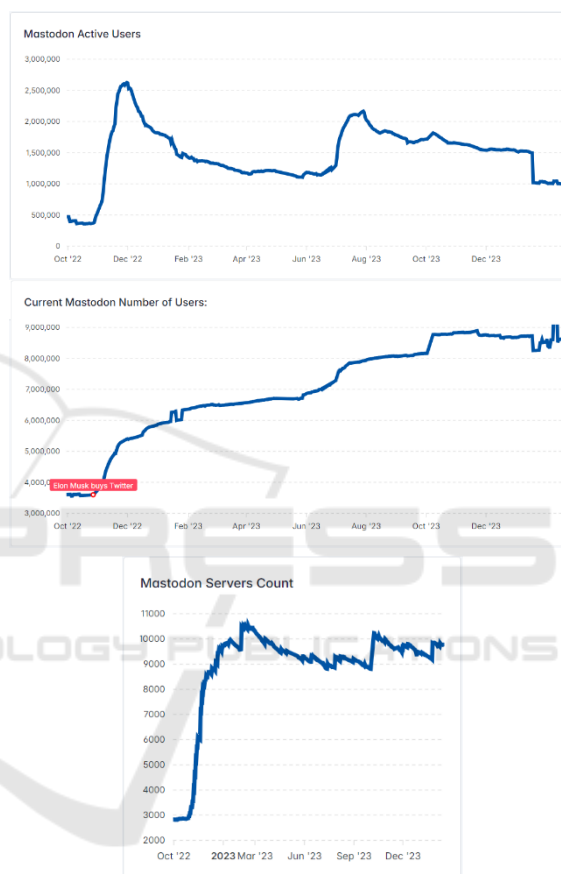


Figure 1: Number of Mastodon servers, users, and active users. Images captured from interactive graphics at <https://mastodon-analytics.com/> (Khun, 2024).

The Guardian article (Nicholas, 2023) provides a comprehensive timeline of how the Mastodon user base evolved in relation to key events involving Twitter. Nevertheless, the challenging process of migrating communities has so far prevented Mastodon from gaining momentum and achieving widespread adoption as a mainstream social media platform. However, even with a significantly smaller user base, with more niche communities compared with the broader audience of Twitter, it is still worth investigating Mastodon as an alternative source of data for social media analysis; that is, given its

considerable advantages, such as higher message character limit, federated architecture, chronological message feed (Lamaj, 2023), and free API access.

An alternative approach to overcome the API limitations newly imposed by social media platforms is to instead collect data by web scraping. However, this approach requires special web tools and add-ons such as BeautifulSoup and Selenium, and the data collection must carefully address legal and ethical concerns (Harrell, 2024).

Social media analysis comprises a large variety of analysis methods, but data is usually sourced from mainstream social media platforms (Zachlod, 2022). Until now, little work has been done on tracking and analyzing data from alternative social media platforms, such as Mastodon. In (David, 2023), the rtoot package is presented, that can be used to collect statuses (a.k.a. toots) from Mastodon and perform some analytics (such as comparing the length of toots from iOS and Web.) In this work, we perform a more thorough examination and investigate if there is value in conducting an analysis of data sourced from Mastodon.social: can tasks such as sentiment, emotion, and topic analysis reveal meaningful trends that are reflective of real-world events?

3 METHODOLOGY

In Figure 2, we show the steps we took to collect Mastodon data and analyze it. The steps are framed in the social media analysis framework presented in (Stieglitz, 2018) (Zachlod, 2022). Our methodological approach follows the steps in (Zachlod, 2022) and (Stieglitz, 2018), therefore, by adequately adjusting the tracking/collection step, the analytical process can be adapted to function with an alternative social media source. We explain each step in detail in the following sections.

3.1 Mastodon API

Mastodon provides access to its data via REST API. We used the Mastodon.py Python wrapper for the Mastodon API to interact with the Mastodon social network. The session.timeline() function was used to collect all messages (called statuses) marked public and whose content string contained one of a set of keywords; Similarly, the session.timeline_hashtag() function was used to collect those statuses marked public and matching one of a set of hashtags. While the account holding the access token for this data collection was hosted on the Mastodon.social instance, we could still access public data originating

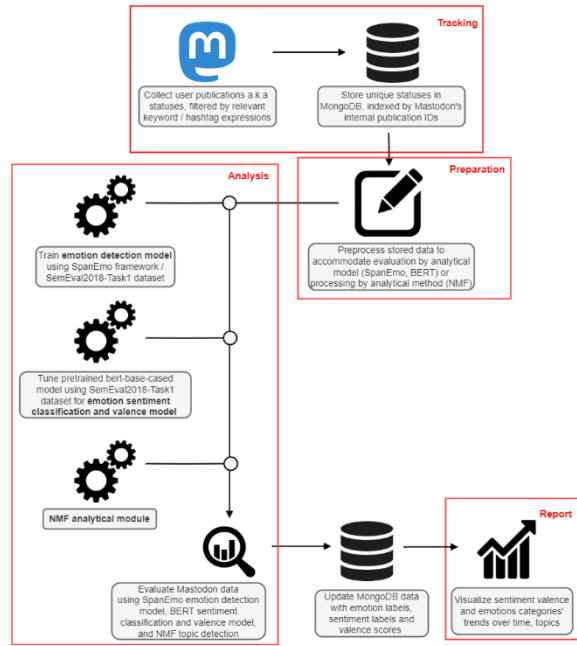


Figure 2: Experimental framework for Mastodon.social data collection and analysis.

from any instance on the federated network. All unique statuses found matching any of the hashtags or keywords were collected and stored in a mongoDB collection for analysis. We chose to focus on two distinct topics: the 2024 US election and competing social media platforms. Each topic was tracked for 7-14 days, and we collected 20,064 social media related statuses, and 6,904 US election related statuses. Table 1 shows the keywords and hashtags that we used for tracking matching statuses for the two selected topics:

Table 1: Keywords and hashtags used for data tracking on Mastodon.social.

Topic	Hashtag list	Keyword list
Social Media	BlueSky, JackDorsey, Facebook, MarkZuckerberg, Meta, Threads, Twitter, ElonMusk, Musk, TwitterMigration, TwitterExodus, X, Xodus, Truth, TruthSocial	Jack Dorsey, Mark Zuckerberg, Zukerberg, Elon, Musk, Elon Musk, BlueSky, Facebook, Meta, Threads, Twitter, X, Truth Social
US Election	POTUS, President, Biden, Democrats, Election2024, Trump, Republicans, USelection, Vote, VoteBlue, VoteBlue2024, Voting	

3.2 Analytical Methods

We conducted several types of analysis: sentiment class and valence prediction, emotion analysis, and topic discovery.

Sentiment classification is a type of analysis where each message is predicted to belong to one of several predefined classes (positive, neutral, negative), based on its content. Similarly, **sentiment valence** prediction associates to each message a numerical score from a range (such as [-1,1]), where the lower the score, the more negative the message is, and the higher the score, the more positive the message is; scores around 0 indicated a neutral or mixed emotional state in the text. Both types of tasks can be approached with a variety of methods (such as VADER (Hutto, 2014), linear regression etc.); more recently, methods based on LLMs have been used for this purpose. We utilized for both tasks a pretrained BERT model with three sentiment classes (negative, positive, and neutral) (Devlin, 2019) (Rathi 2020) that we tuned on a dataset from the SemEval2018-Task1 (Mohammad, 2018). Specifically, we used the tweet set combined from the 2018-Valence-oc-En-train.txt and 2018-Valence-oc-En-dev.txt files, where messages with “Intensity Class” equal to -3, -2, or -1 were assigned to the “negative” class, messages with “Intensity Class” equal to 1, 2, or 3 were assigned to the “positive” class, and the “neutral” class consisted of all messages with “Intensity Class” equal to 0.

The tuned BERT model was then used to predict sentiment class labels and valences for the statuses in our US election and social media Mastodon datasets.

Emotion prediction is tasked with determining which emotions from a given set are present in a message. We followed the approach from SpanEmo (Alhuzali, 2021) (Alhuzali, 2021a) and trained a SpanEmo model on the data from 2018-E-c-En-train.txt and with validation data the 2018-E-c-En-dev.txt (Mohammad, 2018). The SpanEmo model obtained was used to predict the expression of anger, anticipation, disgust, fear, joy, love, optimism, hopeless [sic], sadness, surprise, and trust emotions in US election and social media Mastodon datasets.

Topic analysis attempts to identify topics or themes in a collection of texts. We used non-negative matrix factorization for identifying topics in our two data collections (Greene, 2017).

All of these analytical methods have been tested for accuracy with good results in other works (such as (Alhuzali, 2021)), and we will not include metrics to reflect their validity in this paper.

What is shown is the daily number of unique statuses as reflected in each sentiment class or each emotion category. The sentiment classes are disjoint, i.e. each message belongs exclusively to one sentiment class. The emotion classes overlap, i.e. each status may display several emotions.

As seen in Figures 3 and 4, there are several peaks and valleys in the sentiment and emotion graphs for the two datasets. For example, anger and disgust emotions peaked in the Social Media dataset on March 5, 2024. That is reflected in the following messages shown in Table 2, which indicate users’ reactions to an ongoing service outage at Meta that impacted Facebook, Instagram, and Threads, among various other microservices.

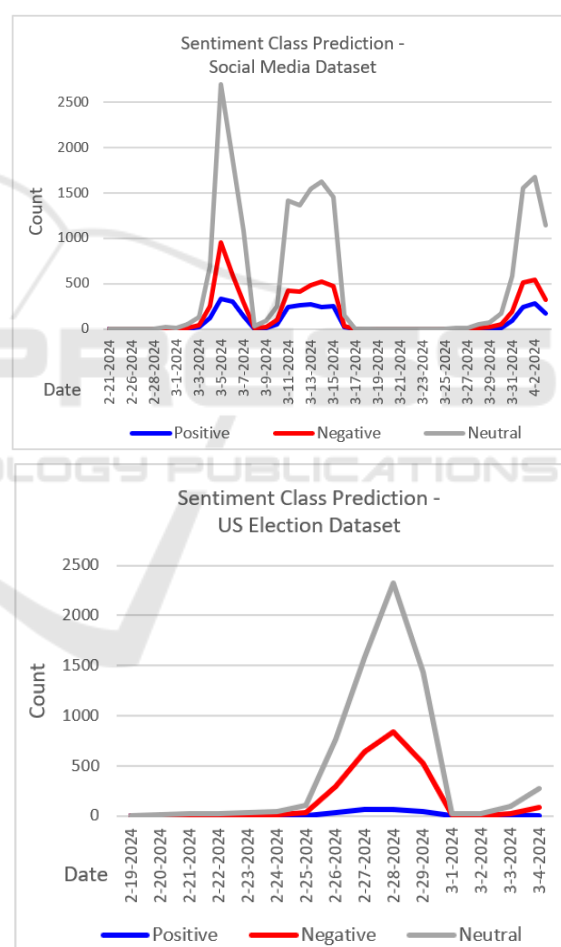


Figure 3: Sentiment classes, daily counts.

4 EXPERIMENTAL RESULTS

Figures 3 and 4 show the sentiment classes and the emotion classes respectively, over time, for the periods during which we collected the respective

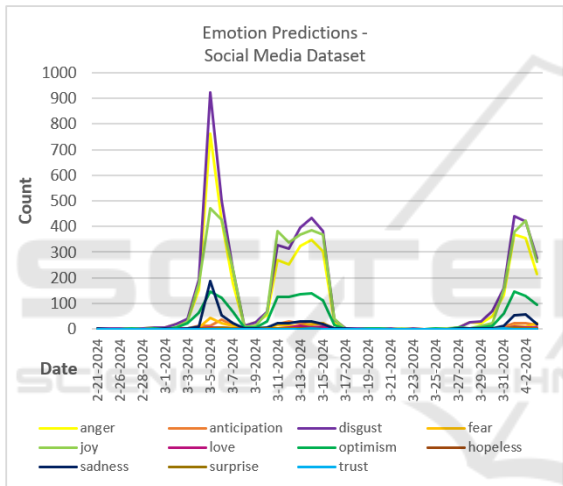
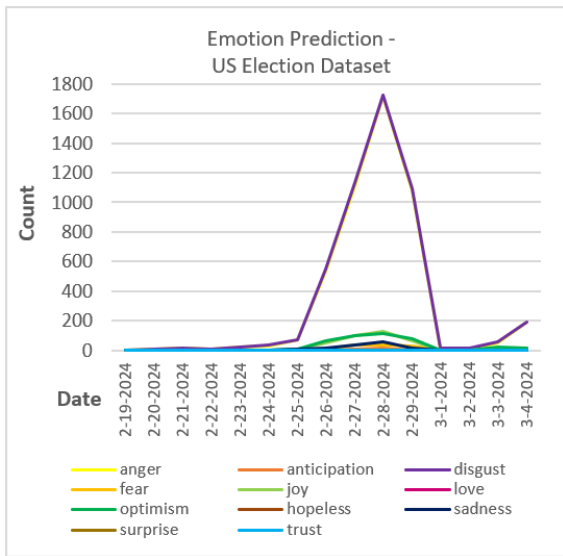


Figure 4: Emotion classes, daily counts.

Table 2: Messages during 2024-03-05 peak.

Date	Message Content	Sentiment (predicted)	Emotions (predicted)
24-03-05 T15:30:56	“Bloody timing of #Facebook going down just after I'd replied to someone in a private message that I've not spoken to for ages instantly making me think I'd fallen victim to some *hack.”	Negative	Anger, disgust
24-03-05 T15:33:03	“MAJOR outage at Meta at the moment. Got booted out of my Facebook account, can't login. Instagram seems to be similarly affected. #Facebook #Instagram #Meta #Outage”	Negative	Anger, disgust, sadness
24-03-05 T15:45:25	“Did Elon buy Facebook?”	Neutral	(null)

We also looked at how sentiment and emotion classes overlapped – which, to our knowledge, has not been investigated before. By verifying that each sentiment class maps into *expected* emotion classes also proves the validity of the independent methods applied for sentiment detection (BERT) and emotion detection (SpanEmo.) For example, anger, disgust, and fear are reasonably associated with negative sentiments; whereas joy, love, and optimism are associated with positive sentiments. Tables 3 and 4 illustrate how the different emotion classes overlap with the negative, positive, and neutral (largely irrelevant and ignored) classes. We highlighted the significant majority sentiment class for each emotion, and we can see that each emotion has highest overlap with the expected class between positive and negative classes:

Table 3: Sentiment and emotion classes overlap in the Social Media dataset.

BERT label	Negative	Neutral	Positive
anger	2348	2045	50
anticipation	6	133	132
disgust	2735	2632	72
fear	116	99	7
joy	57	2494	2053
love	2	23	60
optimism	37	771	784
hopeless	4	0	0
sadness	433	134	7
surprise	10	12	6

Table 4: Sentiment and emotion classes overlap in the US Election dataset.

BERT label	Negative	Neutral	Positive
anger	2189	2708	23
anticipation	2	30	7
disgust	2236	2736	20
fear	95	35	0
joy	11	253	148
love	0	3	3
optimism	7	285	124
hopeless	1	0	0
sadness	95	46	0
surprise	3	12	0

Figure 5 shows the daily average sentiment value for the observed datasets, in the observed time windows. The values are negative for all days in the US Elections dataset, and mostly negative or neutral (around 0) for the Social Media dataset. Again, peaks and valleys are noticeable, but not all are significant – that is because some of these points represent a very small number of messages; for example, the most negative point in the Social Media dataset is recorded

for March 23rd, when there were only 4 statuses collected on the topic, 1 neutral, and 3 negative. On March 5th, when there was a peak of positive and negative counts, the overall average valence shown is -0.11, since the valences of the statuses in two sets, the positive and negative, largely cancel each other out. So, the volume of messages that contribute to the average should be taken into consideration when the average sentiment valence is interpreted.

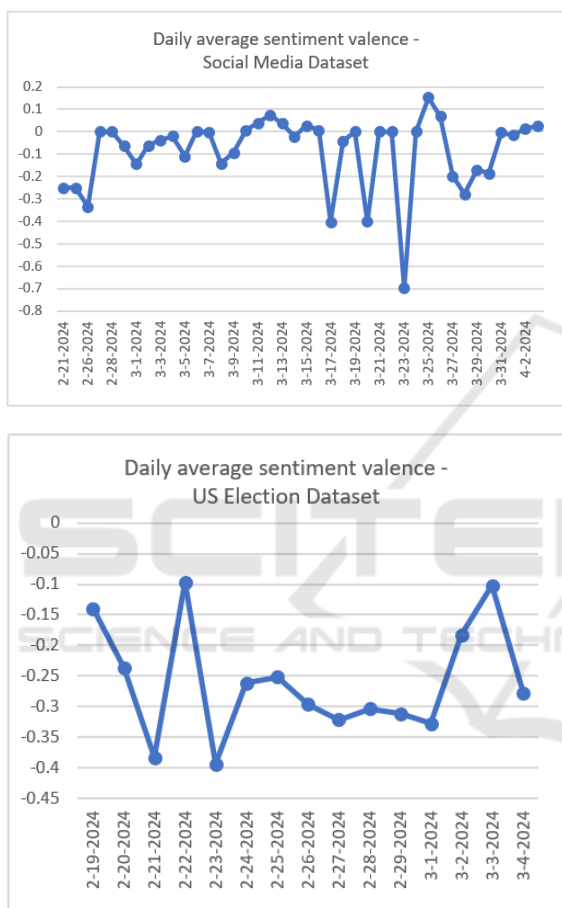


Figure 5: Sentiment valence, daily averages.

Finally, we performed topic analysis using non-negative matrix factorization (NMF) (Greene, 2017), (NMF documentation for scikit-learn, 2024) on a subset of 1755 statuses from the Social Media dataset for March 5, 2024 (the day with the maximum volume of messages in the observed interval), and which had the language specified as English (although some of these were actually in a different language). Figure 6 shows a t-SNE projection representing the words similarities obtained during topic analysis for this dataset:

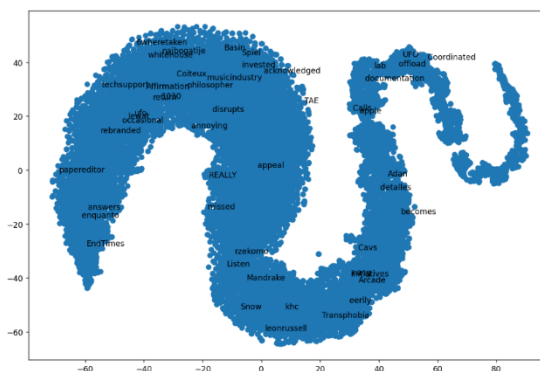


Figure 6: Words in Word2Vec model, t-SNE projection.

The representative terms determined for the NMF model; four topics are shown below. The number of topics was selected based on the coherence and separation of the topic terms:

- Topic 1: `https`, `com`, `news`, `score`, `id`, `item`, `ycombinator`, `url`, `title`, `discussion`
- Topic 2: `value`, `form`, `labour`, `marx`, `exchange`, `like`, `one`, `production`, `bailey`, `commodity`
- Topic 3: `facebook`, `instagram`, `meta`, `outage`, `com`, `threads`, `www`, `https`, `2024`, `users`
- Topic 4: `bluesky`, `network`, `also`, `nbsp`, `fediverse`, `people`, `data`, `app`, `website`, `protocol`

Topic 3 reflects the Facebook outage that occurred on March 5, 2024.

5 CONCLUSIONS AND FUTURE WORK

Our experimental results show that there is value in analyzing data extracted from new and alternative social media platforms such as Mastodon. Analytical tasks such as emotion, sentiment, topic analysis can still reveal trends and identify reflections of real-world events.

In the future, we want to observe and compare various social media platforms for data completeness, quality, volume. We want to investigate how data collected from Mastodon, Threads, BlueSky, and other federated (or soon-to-be federated) platforms compares with that from the mainstream social networks in terms of discourse, sentiment, topics etc. – are they similar or very different? We also plan to research the difference in the data value depending on the collection method (RSS, API, HTML scraping.)

REFERENCES

- Zachlod, C., Samuel, O., Ochsner, A., Werthmüller, S. (2022). Analytics of social media data – State of characteristics and application. In *Journal of Business Research*, Vol. 144, May 2022, pp. 1064-1076.
- Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. In *International Journal of Information Management*, Vol. 39, April 2018, pp.156-168.
- Mastodon statistics (2024), online at <https://api.joinmastodon.org/statistics>, visited July 2024.
- Khun, E. (2024) - @erickhun@mastodon.social, Mastodon growth dashboard, online at <https://mastodon-analytics.com/>, visited July 2024.
- Nicholas, J. (2023). Elon Musk drove more than a million people to Mastodon – but many aren't sticking around, online at <https://www.theguardian.com/news/datablog/2023/jan/08/elon-musk-drove-more-than-a-million-people-to-mastodon-but-many-arent-sticking-around>, Jan 7, 2023.
- Mastodon.social (2024), online at <https://mastodon.social/about>, July 2024.
- Lamaj, D. (2023), Twitter Vs. Mastodon: Which is a Better Alternative? (Pros and Cons), online at <https://publer.io/blog/twitter-vs-mastodon/>, July 2023.
- Harrell, N.B., Cruickshank, I., Master, A. (2024). Overcoming Social Media API Restrictions: Building an Effective Web Scraper, In *Proceedings of the ICWSM Workshops*, June 2024.
- David Schoch, D., and Chan, C.-H. (2023). Software presentation: Rtoot: Collecting and Analyzing Mastodon Data. In *Sage Journal of Mobile Media & Communication*, Volume 11, Issue 3, 575-578, <https://doi.org/10.1177/20501579231176678>
- Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*. 8 (1).
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, June 2019.
- Rathi, P. (2020). Sentiment Analysis using BERT, code repository, online at <https://www.kaggle.com/code/prakharrathi25/sentiment-analysis-using-bert>.
- Mohammad, S., Bravo-Marquez, F., Salameh, M. and Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June 2018.
- Alhuzali, H. and Ananiadou, S. (2021). SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, April 2021.
- Alhuzali, H., (2021a). SpanEmo, code repository, online at <https://github.com/hasanhuz/SpanEmo>.
- Greene, D. (2017). Topic modelling with Scikit-learn. Presented at PyData Ireland, September 2017, github repository online at: <https://github.com/derekgreene/topic-model-tutorial/>
- NMF documentation for scikit-learn (2024), online at <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>, visited 2024.