PROJECT COURSE REPORT

Social Graph Miner: Climate Discourse Analysis

Damiano Orlandi

January 13, 2026

Contents

1 Introduction

The 21st century is defined by the most advanced phase of globalization and continuous technological improvements. Today, innovations have created a state of constant interconnection among worldwide citizens: news from distant countries, cultural trends, and social debates circulate globally in real time, profoundly shaping our daily lives [16]. Yet, while this hyper-connected environment offers opportunities, it also brings new challenges. We, as humans, increasingly adapt our behaviors, opinions, and even our identities to the systems that govern our digital experience, often unconsciously and sometimes at the expense of our autonomy.

In this landscape, actively choosing how and where to engage online has become an essential skill. Rather than passively accepting the influence of platforms, individuals now seek to shape their own digital environments. The phenomenon known as "X-migration" is a clear example of this trend: users, motivated by concerns over platform management, algorithmic control, or corporate ethics (such as those triggered by changes under Elon Musk's guide), deliberately move away from mainstream platforms in favor of substitute [17].

Inspired by these large-scale acts, it came natural to delve into all the possible alternatives that worldwide citizens have when they want to build up a digital identity, that are summarized into the concept of Fediverse. Such term represents the set of federated social media that offers an alternative to centralised, and non-interoperable platforms through open protocols and policies [15]. Microblogging, social networking and content sharing are some of the services provided by the softwares built on top of such architecture. Among the main programs, Bluesky, a new social media publicly released in February 2024, will be analyzed in this project.

Moved by the interest of properly deepening such new social platform I decided to merge project course, internship and thesis to conduct a wide analysis and to narrow it down to one of the most spoken theme of the last decade: climate change. Specifically, the research is oriented in the comprehension of the climate discourse by using the latest approaches of the computational social science research field. The time span between February 2024 and July 2025 will be taken into account to reveal main and subtle topics of the climate realm and to analyze the way users bound with each other thanks to the topics they deal with. Therefore advanced topic modeling techniques and network science paradigms will be applied to study both textual and graph data. The lack of comprehensive, publicly available datasets for such platform required the creation of a custom data collection system and, therefore, to address it, I invested the project course to develop a "social graph mining system" designed to extract and store millions of posts via public APIs, forming the foundation of a wider analysis.

The report describes the procedure, including NLP methodology and tools, applied to to build such a system in a comprehensive way.

1.1 The innovative social media

Bluesky is a microblogging social media service publicly accessible from the 6th of February 2024 with a story rooting back five years before. It was 2019 when Jack Dorsey, the former Ceo of Twitter, firstly had the idea to spend some energies in developing a way to create a decentralized and open social media. He started from the platform he founded and tried to improve it. The way he chose was to reduce the centralized content moderation and create a new protocol capable of leaving a wider data control to the users. In 2022, after the nomination of a new CEO, Jay Graber, and the acquisition of Twitter by Elon

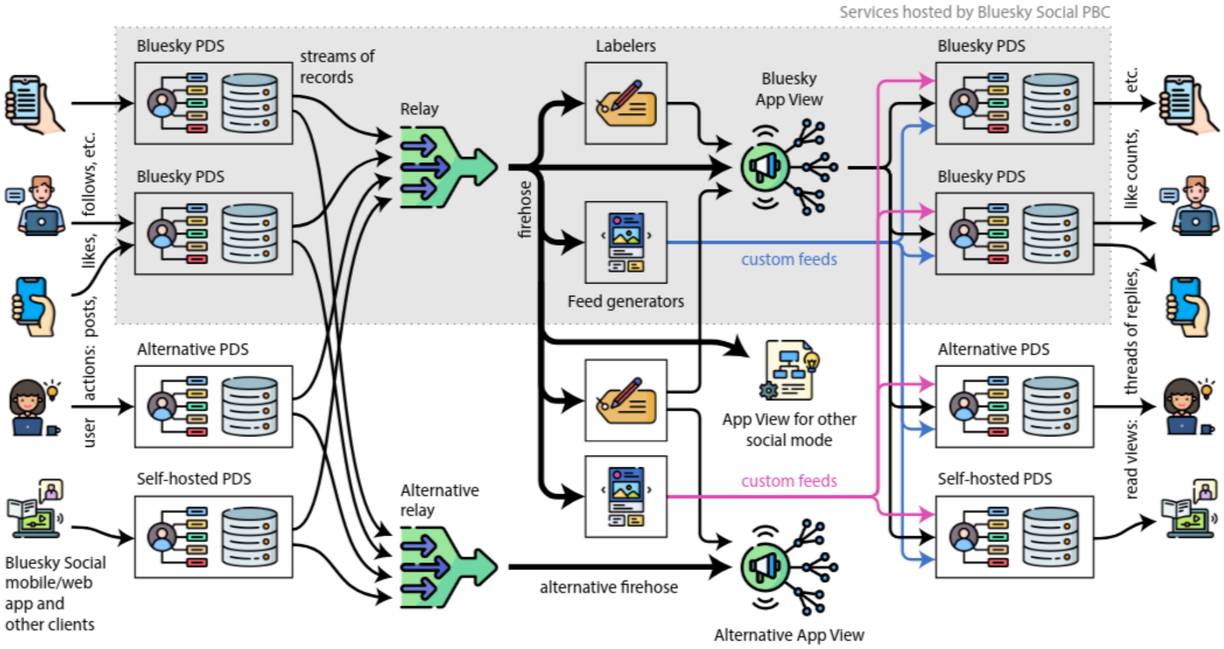

Figure 1: The main services involved in providing Bluesky, and data flows between them. Icons from Flaticon.com.

Musk, Bluesky turned into a Public Benefit LLC and established the basis to the invite-only IOS-beta launched in 2023. The first year, it reached roughly 1 million of users and after public sign-ups opened in 2024, its user base shifter to about four million profiles within the first week. This trend highlights the strong curiosity and clear demand for an alternative social platform. The whole social media holds on a new protocol developed ad hoc Atproto. Standing for Authenticated Transfer Protocol, it is Blueksy core [13] and, as other federated protocols, instead of a single operator owning identity and data, it decomposes the stack into interoperable services so users can switch providers without losing identity or social graph.

Atproto achieves this by anchoring everything to a cryptographic user repository hosted on a Personal Data Server (PDS). Each account' posts and actions are stored as typed records (defined by shared Lexicon schemas) inside a signed repository structure, so replicas and indexers can verify integrity instead of blindly trusting a single database. Identity is split into a human-friendly handle (DNS-style, often a domain you control) and a stable DID (cryptographic identifier), where the handle resolves to the DID via a DNS TXT record or an HTTPS endpoint, enabling domain-based self-verification. For scale, many PDS publish changes outward: a Relay aggregates updates into a real-time firehose stream, and an AppView consumes that stream. On top of the same shared data plane, specialized services remain pluggable: feed generators can compute alternative rankings/feeds, and labelers can attach moderation labels that clients may subscribe to and enforce. Interoperability is kept strict by exposing everything through XRPC APIs whose request/response shapes are standardized via Lexicon, letting different clients and servers implement the same protocol surface. The image above shows its core structure (figure 1).

2 Methodology

The **social graph mining system** developed in this project is a modular data engineering framework that leverages public APIs to automatically collect, preprocess, and store large volumes of social data in a normalized format suitable for downstream analysis.

The methodological workflow is structured in four main phases:

1. Identification of PDS domains;
2. Definition of filtering criteria for targeted data collection;
3. Survey of platform-specific APIs and authentication procedures;
4. Construction of a scalable architecture to extract and store all data on Google Cloud.

2.1 PDS Domain

Considering the segmented nature of the Fediverse, the first needed step is to identify those PDS domains interpolating the majority of data flow as they will be used in the next step to call the Bluesky API. Accordingly to the social media documentation, the only way to verify the real volume of all possible PDS hosting domains, would have been to open a link to the firehose in February 2024 and store all domains used until July 2025. Unfortunately, such approach is unfeasible as the research started after Bluesky opened the sign-ups publicly. However, as stated in [2], whenever a user creates a profile on Bluesky, the call to `com.atproto.server.createAccount` is sent to the PDS Entryway (that is a PDS orchestrator managing all PDS), therefore it is chosen to use `bsky.social` as main domain.

2.2 Filtering Words

A crucial aspect of the data extraction process is the definition and application of filters to isolate climate change related content from noise. The core methodological challenge is to avoid bias while ensuring that the dataset is both representative and relevant. The applied procedue can be briefly summarized in three steps:

- Two dataset are queried by using (`"climate change"`) and `"#climatechange"` to assess which strategy put in place for the final dataset extraction. This approach is essential due to the nature of the content inside the posts and the way the API is written (the query can be both an hashtag or a normal text).
- A baseline dataset is retrieved to investigate the normal presence of hashtags climate-related in random posts, by using a random selection of stopwords from nltk Python library.
- Idenitification of the most meaningful hashtags that will be implemented for the final data extraction.

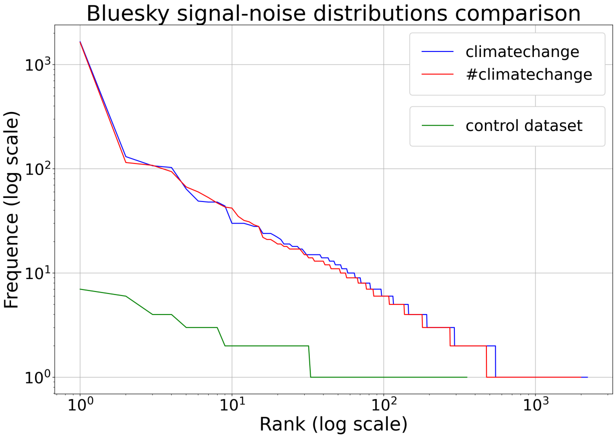

Figure 2: Distribution of hashtags in random and climate-filtered samples.

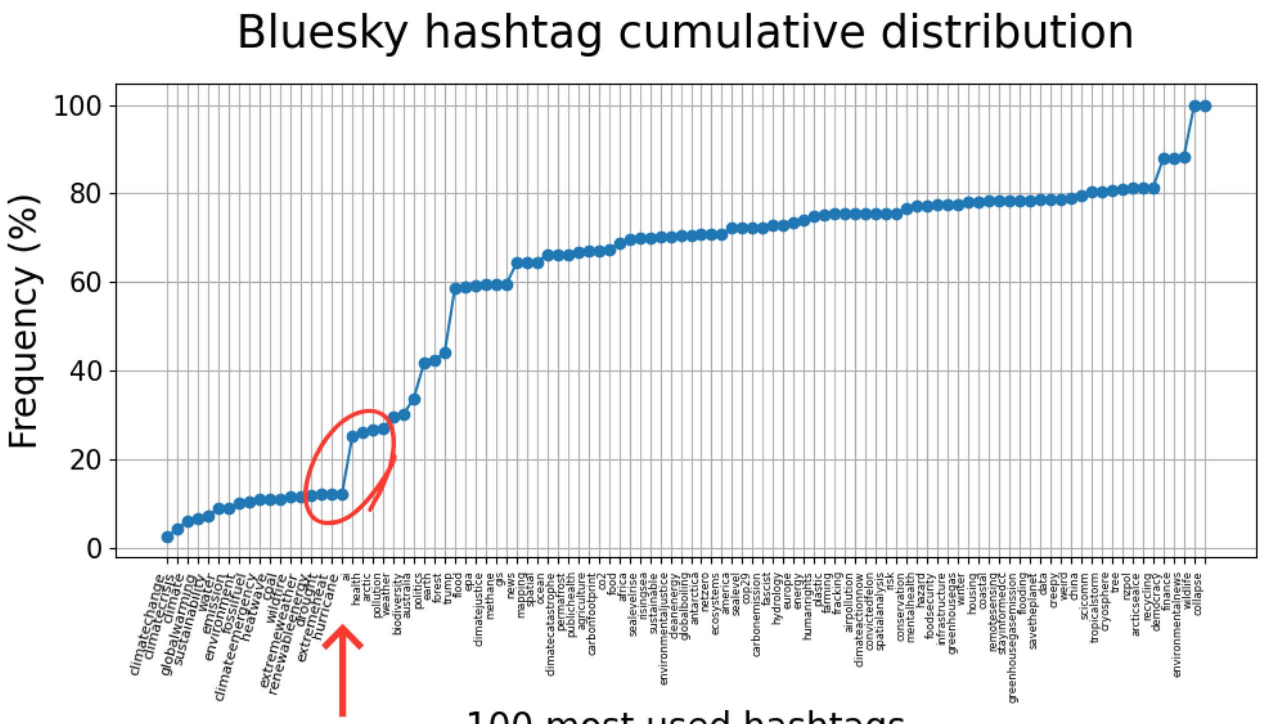

Figure 3: Cumulative distribution of hashtag frequencies in a daily sample.

Random Sampling and Baseline Distributions To construct an unbiased baseline, a random sampling procedure is followed: for each month in the analysis window, 10 random hours per day are generated, posts matching the primary filtering criterion (`"climatechange"`) are queried, duplicates are discarderded, English-language content is filtered, and 100 posts per month are selected. However, due to Bluesky API restrictions, a dual approach is used: both `"#climatechange"` and `"climatechange"` (no hashtag) queries are used as filters and their distributions are computed to assess potential differences. Since no significant variation in hashtag distributions is observed (Figure 2), the analysis is conducted with the simpler `"climatechange"` (not including the hashtag) dataset.

Moreover, to avoid any unbalanced search, another dataset is retrieved by querying a random list of stopwords (from nltk library) - in the same way previously described. The reason why it is critical to estimate the "background" distribution of hashtags and terms is for distinguishing signal (climate discourse) from noise. The result, as shown in Figure 2, is a heavy-tailed distribution with climate-related hashtags underrepresented in random samples.

Frequency Analysis and Thresholding Computing z-scores (relative frequencies of hashtags) was found to be uninformative, given the sparseness of climate hashtags in random samples and the natural skew of hashtag frequencies. Percentile cutoffs (e.g., 99.5th percentile) were considered, but did not materially improve the separation between noise and relevant hashtags. Instead, a frequency-based approach was adopted: extracting the 100 most frequent hashtags from the dataset queried by `"climatechange"` and then computing the cumulative frequency distribution for all hashtags on a randomly selected day enables a robust identification of dominant terms without over-representing rare hashtags. (Figure 3) shows such trend.

Selection of Hashtags.

The final lists of hashtags used for extraction and analysis are reported below:

- **Bluesky list:**["climatechange", "climatecrisis", "climate", "globalwarming", "sustainability", "water", "emission", "environment", "fossilfuel", "climateemergency", "heatwave", "coal", "wildfire", "extremeweather", "renewableenergy", "drought", "extremeheat", "hurricane"]

The choice has been made due to the distribution spike (red circled on the plot) and its semantic meaning transposed in the real world: it is clear that the term "ai" is out of context compared to the 18 previous words. All of the terms cover subsets related to climate narrative, both from a general perspective and a specific one (e.g. "drought", "hurricane", "coal", "wildfire" etc.) leading to a potential extensive search, both fine-grained and wide. Even though some of them may represent broad topics and invade other public debates, like "water", it is still meaningful to include them in the analysis as they may be useful to identify some subtle subtopics of the climate narrative.

2.3 APIs

The chosen API is `https://bsky.socialxrpc/app.bsky.feed.searchPosts` as it is the only one that grants the opportunity to apply both hashtag and temporal filters to the query. Such function is receiving as input the filtering hashtag list (used both in the preliminary step of searching "climatechange" and in the proper data extraction with the list of hashtags of the distribution), the language (set as "en", to restrict the analysis to only english speakers), the domain ("bsky.social"), the limit of retrieavable posts per call (set to 100, maximum) and the cursor for pagination. However, given the structure of the Atproto and given that the calls are made to the AppView endpoint, multiple domains are retrieved during the search. Figure 4 shows the distribution of the most common terms obtained after the "Frequency Analysis and Thresholding" step (section 2.2), and it confirms what stated in 2.1 ("bsky.social" is the most used domain). An authentication token as been generated on the Blueksy personal profile in the software developer section as required by the platform documentation to have access to PDS entryway data. Such information can be found here: `https://docs.bsky.app/docs/api`.

3 Architecture

All data extraction was performed on the university high performance computing infrastructure in order to have scalable access to computational resources and network bandwidth. To ensure reproducibility and portability, the entire software stack was encapsulated within a containerized environment (Docker) based on Ubuntu (CPU-only), including all required Python libraries and system dependencies. The container was deployed across the HPC cluster to leverage multiple compute nodes, enabling large-scale parallel data retrieval from the Bluesky AppView API. This is the reason why the extraction pipeline followed a distributed

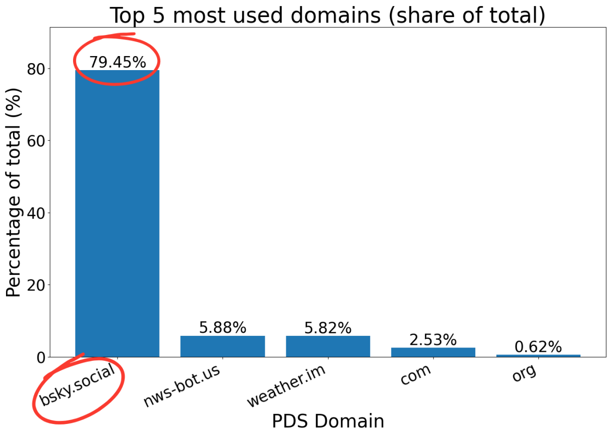

Figure 4: Distribution of the most common domains after the hashtag filtering step

execution model based on Dask, where one node was used as a lightweight scheduler, while multiple nodes acted as workers. Each worker executes between four and five independent processes to collect data by calling Bluesky AppView. To avoid authentication issue and to simplify data partitioning, each worker was assigned a unique API token and was responsible for a single month of data. Before starting the retrieval process, workers are connected to a shared Google Cloud Storage bucket on which data are moved (from the local storage) in batches of 5,000 records, reducing I/O overhead and preventing bottlenecks. The final obtained setting is a folder in the cloud with as many subfolders as the number of months (16) containing one JSONL file with all the posts information.

3.1 Structure

Dask is a Python-native distributed computing framework that perfectly adapts to parallel and cluster-scale environments. It represents computations as directed acyclic graphs (DAGs) of fine grained tasks, which are dynamically scheduled and executed across multiple workers by a centralized scheduler. This design allows Dask to efficiently manage task dependencies, balance workloads, and recover from partial failures without requiring tight coupling between tasks. It supports asynchronous control over resource allocation, and deployment on HPC clusters. In this project, Dask enabled the coordination of thousands of independent API calls across multiple nodes while maintaining control over concurrency, memory usage, and task isolation.

3.2 Cloud storage

Google Cloud Storage (GCS) is an object-storage service where data are stored as immutable objects inside a bucket (a globally unique namespace). Instead of a filesystem with directories, GCS relies on object keys and is optimized for high durability, high availability, and massive horizontal scalability. This makes it a strong fit for distributed pipelines, because many Dask workers can upload objects concurrently to the same bucket without needing a shared POSIX filesystem or cross-node locking.

In this project, data were persisted to GCS using a dedicated service account. Each worker authenticated through Application Default Credentials backed by a service-account key, then wrote JSONL data to a shared bucket, matching the temporal partitioning of the extraction tasks. Posts were accumulated locally and uploaded in batched writes to reduce per-request overhead and mitigate I/O bottlenecks under parallelism. Overall, this setup enabled robust, concurrent ingestion from multiple HPC nodes while keeping storage operations simple, scalable, and reproducible.

3.3 Data extraction

Multiple approaches were tried, but the most convenient and final one was to assign one week retrieval per process of each worker and iterate over the hashtag list. Due to the impossibility of retrieving the full outer join among tags, it was inevitable to implement a temporal set (Python object) on which store all possible posts and move to the cloud bucket only the unique ones. The removal of duplicates was achieved both by the inner characteristics of the python set and by a second check. This step is known to reduce the speed and introduce a potential memory issue, but was needed given the API structure and the importance of store unique posts.

Token expiration management Even if each worker was initially assigned credentials, a token management system was needed to control

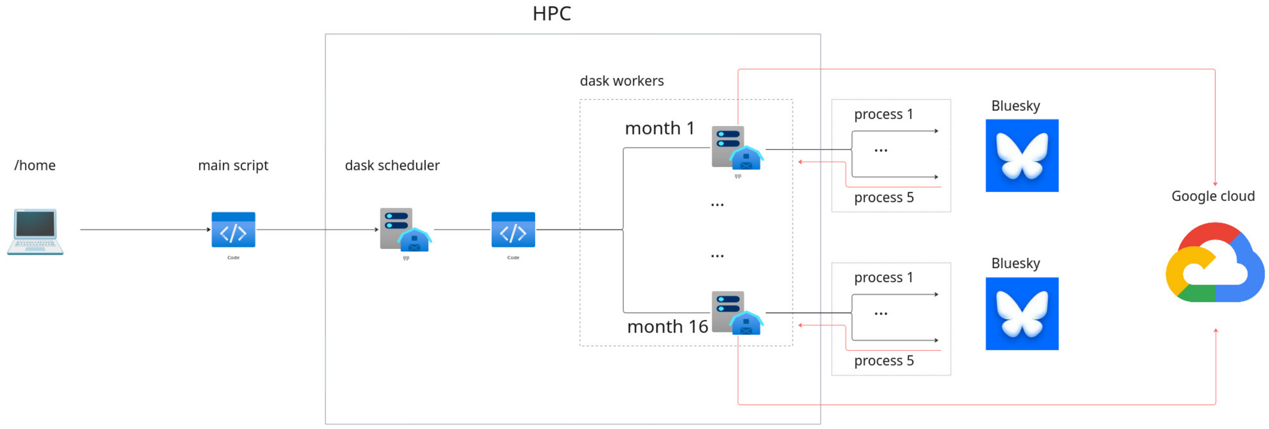

Figure 5: Data extraction workflow

the token expiration. Token validity lasts 120 minutes and its renewal is possible by opening a new session. Therefore, by tracing the time of the first call and obtaining a new authentication key before expiration, it was possible to keep extracting without limits. The token manager calls `server.createSession` to obtain both `accessJwt` and `refreshJwt`, so that is possible to set `refreshSession` as `True` and keep the session active.

Fault tolerance An error handling "system" was thought to manage potential server, time or authentication issues. Thanks to a few functions, it was possible to avoid any local or remote server issue by retrying the calls up to a fixed amount of time; bypass rate limits by automatically waiting until the 600 seconds were passed or reset the Jwt calls to get new credentials. As before, this approach is the result of multiple attempts empirically proven.

API limits A barrier faced during extraction is the API limit, set by the provider. The threshold defined by the platform is 3000 calls per 5 minutes, leading to a maximum amount of retrievable posts per day (without any error) equals to 864000 (where "limit" is intended to be per token).

3.4 Data type

Bluesky (AT Protocol) returns a post view with top-level `uri`/`cid`, `author`, and the actual post in `record` (`text`, `langs`, `facets`, `createdAt`) as shown below:

`Field`	**Meaning (ingestion note)**
`uri`	AT URI (global post ID)
`cid`	Content ID (CID)
`record.createdAt`	ISO-8601 timestamp
`record.text`	Plain text
`record.langs[]`	Languages (array)
`author.did`	Author DID (stable)
`author.handle`	Author handle (domain-qualified)
`author.displayName`	Author display name
`features[].tag`	Hashtags (facet #tag)
`features[].mention`	Mentions (if present)
`features[].uri`	Richtext links
`embed.images[]`	Image embeds (if any)
`parent.uri`	Parent post (if reply/thread)
`replyCount, repostCount, likeCount, quoteCount`	Engagement counters
`labels[], viewer.embeddingDisabled`	

4 Conclusion

This data engineering system was an essential foundation for the thesis, because it enabled the study of climate-change discourse at a scale. By providing a reliable way to gather and consolidate content from Bluesky into a unified, queryable dataset, the pipeline transforms dispersed data into a coherent research resource. Its emphasis on reproducibility, scalability, and consistent data organization ensures that downstream analyses can be carried out efficiently. In this sense, the contribution is not only technical, it establishes the methodological infrastructure required to support rigorous computational social science results throughout the thesis. With the dataset now in place, the next phases will focus on the following analytical tasks:

- **Topic modelling** Textual data will be studied to detect the most treated topics, their evolution, and stabilization over time. A modular approach (similar to BERTopic) will be utilized to reduce embed posts content, reduce its dimensionality and cluster the resulting representation in coherent and semantic concepts.
- **Network analysis** Interaction networks will be extracted to build a multi-layer representation of the discourse and its dynamics. Graphs based on follower relations, reposts, and mentions will be analysed to identify community structures and to quantify the strength and patterns of connections between users.

These steps will enable a deep understanding of how climate change narratives form, spread, and stabilise — and how user interactions influence specific topics within the discourse.

References

[1] AT Protocol. At protocol. https://atproto.com/.

[2] Bluesky. Api documentation. https://docs.bsky.app/docs/advanced-guides/entryway, .

[3] Bluesky. Api documentation. https://docs.bsky.app/docs/api, .

[4] Bluesky. Api documentation. https://docs.bsky.app/docs/category/http-reference, .

[5] Bluesky. Api documentation. https://docs.bsky.app/docs/api/app-bsky-feed-search-posts, .

[6] Biraj Dahal, Sathish A. P. Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):24, 2019. doi: 10.1007/s13278-019-0568-8. URL https://doi.org/10.1007/s13278-019-0568-8.

[7] Dask. Dask documentation. https://docs.dask.org/en/stable/.

[8] Ramit Debnath, Ronita Bardhan, Darshil U. Shah, Kamiar Mohaddes, Michael H. Ramage, R. Michael Alvarez, and Benjamin K. Sovacool. Social media enables people-centric climate action in the hard-to-decarbonise building sector. *Scientific Reports*, 12:19017, 2022. doi: 10.1038/s41598-022-23624-9. URL https://doi.org/10.1038/s41598-022-23624-9.

[9] Thierry Declerck and Piroska Lendvai. Processing and normalizing hashtags. In *Recent Advances in Natural Language Processing*, 2015. URL https://api.semanticscholar.org/CorpusID:14353321.

[10] Dimitrios Effrosynidis, Alexandros I. Karasakalidis, Georgios Sylaios, and Avi Arampatzis. The climate change twitter dataset. *Expert Systems with Applications*, 204:117541, 2022. doi: 10.1016/j.eswa.2022.117541. URL https://doi.org/10.1016/j.eswa.2022.117541.

[11] A. Galdeman and L. M. Aiello. Mapping the climate change landscape on tiktok. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2614–2621, 2025. doi: 10.1609/icwsm.v19i1.35962. URL https://doi.org/10.1609/icwsm.v19i1.35962.

[12] Google Cloud. Cloud storage documentation. https://cloud.google.com/storage/docs/introduction.

[13] Martin Kleppmann, Paul Frazee, Jake Gold, Jay Graber, Daniel Holmgren, Devin Ivy, Jeromy Johnson, Bryan Newbold, and Jaz Volpert. Bluesky and the at protocol: Usable decentralized social media. In *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet*, DIN '24, page 1–7, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712524. doi: 10.1145/3694809.3700740. URL https://doi.org/10.1145/3694809.3700740.

[14] Arianna Pera and Luca Maria Aiello. Shifting climates: Climate change communication from youtube to tiktok. In *ACM Web Science Conference (WebSci '24)*, pages 1–6, 2024. doi: 10.1145/3614419.3644024. URL https://doi.org/10.1145/3614419.3644024.

[15] Robert Riemann. Federated social media platforms. TechDispatch QT-AD-22-001-EN-NI, European Data Protection Supervisor (EDPS), July 2022. TechDispatch #1/2022.

[16] G. D. Stasberger. Media globalization: Connecting the world through information and culture. *Global Media Journal*, 21(64), 2023. doi: 10.36648/1550-7521.21.64.387. Published 23 Aug 2023. Accessed 12 January 2026.

[17] Adam Volle. fediverse. Encyclopedia Britannica, June 2025. URL https://www.britannica.com/technology/fediverse. Accessed 12 January 2026.