Sentiment Analysis with Large Language Models on Bluesky: Tag Groupings and Decentralized Social Media

Olga Grigorieva, Małgorzata Kurcjusz-Gzowska, Elen Muradyan, Suren Mnatsakanyan
Warsaw University of Technology
olga.grigorieva.stud@pw.edu.pl, malgorzata.kurcjusz.stud@pw.edu.pl,
elen.muradyan.stud@pw.edu.pl, suren.mnatsakanyan.stud@pw.edu.pl

Supervisor: Anna Wróblewska
anna.wroblewska1@pw.edu.pl

Abstract

Large language models can be used for sentiment analysis on social media, but the research usually focuses on centralized platforms, for example Twitter and Facebook. Bluesky is built on the decentralized AT Protocol, introducing new challenges since moderation can move across services, users create their own feeds, and tagging practices are still evolving. These features require revised methods for generating and using sentiment signals. Our project proposes a systematic study of LLM-based sentiment analysis for Bluesky. We will (i) explore existing Bluesky-specific sentiment datasets for text, use existing datasets where available (if these datasets are not sufficient, we will acquire additional data from the Bluesky firehose). (ii) We will evaluate LLM and transformer baselines, and (iii) develop methods that combine LLM sentiment with hashtag clustering and graph-based representations suited to decentralization. The project will make open-access datasets, models, and evaluations. Thanks to this, Bluesky can become a reference point for sentiment analysis on decentralized platforms.

Keywords: large language models, sentiment analysis, decentralized social media, Bluesky, AT Protocol, hashtag grouping, hashtag clustering, multimodal sentiment, annotation practices, bias and fairness.

In this paper, we use the terms 'tag' and 'hashtag' interchangeably to refer to hashtags used within social media posts.

1 Introduction

Large language models (LLMs) have changed sentiment analysis on social media. Earlier methods relied on lexicons or supervised classifiers for short, noisy texts, while transformers enable context-sensitive representations and domain-specific fine-tuning. GPT-style, LLaMA-family, and other open-source LLMs now deliver strong zero- and few-shot performance across various domains, including news, finance, health, and multilingual social media (Zhang et al., 2024). However, success depends on prompt design, domain alignment, and calibration. At the same time, the architecture of social media itself is shifting. Bluesky and the AT Protocol separate identity, hosting and feed generation, which gives users to move between providers and enables the operation of multiple custom feed generators and labeling services (Kleppmann, 2024). This decentralized design raises new questions: how do sentiment signals behave when feeds, labeling, and moderation are modular, and how does decentralization influence their flow and interpretation? The role of hashtags is also crucial, because they shape discovery, format topics, and help to identify the community. Previous work on centralized platforms has used deep learning and graph-based methods for tag recommendation, dynamic adaptation, and clustering (Djenouri et al., 2019; Liou et al., 2020). And recent studies leverage LLMs to refine topics and explain clusters, but they rather ignore decentralized platforms. Our goal in this project is to integrate these strands by designing and evaluating LLM-based sentiment analysis methods specifically for Bluesky:

- explore existing Bluesky-native sentiment and tag datasets, including the Bluesky Social Dataset and POLITISKY24 (Rostami et al., 2025; Failla et al., 2025), which include user-generated posts, political stance labels,

and, where available, multimodal content.

- benchmark LLMs and standard transformer baselines on decentralized social media data,
- develop LLM-supported tag-grouping methods that fit the AT Protocol architecture;
- assess bias, fairness, and uncertainty when sentiment and tags are used in simulated Bluesky ranking pipelines.

The outcome will be a professional, reproducible framework for studying sentiment on decentralized social networks, along with concrete tools and datasets that other researchers can reuse.

2 Literature Review

2.1 Sentiment Analysis on Social Media

LLMs perform well on standard polarity classification, although dedicated architectures continue to perform better on structured tasks such as aspect-based sentiment analysis and opinion-role extraction (Zhang et al., 2024). Existing benchmarks show us that general-purpose LLMs can compete with fine-tuned transformers, especially when only small labeled datasets are available. Domain-specific work reflects this mixed picture. GPT-style and encoder-decoder models can match or outperform fine-tuned transformers with well-crafted prompts, but their performance drops on noisy or highly specialized material (He et al., 2024). Industry reports highlight the benefits of rapid, multilingual deployment, while also acknowledging ongoing challenges related to prompt design, safety, and operational costs. On platforms more similar to Bluesky, fine-tuned BERT, BERTweet and open LLMs boost political sentiment detection. Recent open models close much of the remaining gap when given enough in-domain data. Predictions are sensitive to linguistic issues such as emojis, sarcasm, code-switching, and non-standard varieties. Paraphrasing noisy posts can raise accuracy. However, it may also erase minority language forms. Multilingual studies show encouraging results with well-written prompts, although performance remains uneven for low-resource languages (Nasution, 2023; Fu, 2023) and this is important for Bluesky, which includes large English and Japanese communities and is becoming more linguistically diverse (Sahneh et al., 2025).

2.2 Decentralized Social Media and Bluesky

Bluesky is built on the AT Protocol. It separates the social graph, identity, and content hosting, allowing providers to interoperate (Kleppmann, 2024). Moderation and feed curation are modular, allowing labeling services and feed generators to run independently. Early analyses of Bluesky's growth point to fast uptake, varied posting patterns, relatively low toxicity, and active moderation, although these studies rely on classical toxicity metrics rather than LLM-based sentiment analysis (Sahneh et al., 2025). Broader research on decentralized protocols shows that decentralization redistributes, but does not eliminate, control over moderation or the structural inequalities tied to it (Huang, 2024).

2.3 Hashtags and Hashtag Groupings

Hashtags help to organise content, support discovery, and influence how topics and forming of communities. Deep learning models outperform bag-of-words methods for predicting hashtags (Djenouri et al., 2019), while approaches such as H-ADAPTS and dynamic graph transformers capture shifting usage patterns and infer new tags (Liou et al., 2020). Co-occurrence graphs and community-detection techniques reveal clusters linked to themes or actors, and LLMs can refine topic labels or reduce noise using clustering tools like BERTopic. Most existing work assumes centralized platforms with stable architectures, leaving hashtag grouping in decentralized, instance-specific environments largely unexamined (Feng et al., 2015).

2.4 Bias, Fairness and Multimodal Sentiment

Bias and fairness are central to LLM sentiment analysis. Fine-tuned models can produce systematic differences across demographic attributes despite achieving high accuracy (Radaideh et al., 2025), whereas adversarial training and post-hoc debiasing reduce bias on Twitter benchmarks (Venugopal et al., 2023). Decentralized networks' normative features (blocklists, opt-in search) may amplify such biases in moderation or ranking (Huang, 2024). Multimodal sentiment adds complexity: images improve classification for short, neutral, or sarcastic text, but current models struggle with cultural references or text-image contradictions and are often poorly calibrated (Jin et al., 2024). Annotation and uncertainty estimation are

critical, as disagreements often reflect ambiguity, and access to images can shift labels (Kadriu et al., 2022). Calibrated models help flag unreliable predictions (Xiao et al., 2023).

3 Research Objectives and Questions

This project aims to develop and evaluate a comprehensive framework for LLM-based sentiment analysis on Bluesky, accounting for tag groupings, multimodality, where applicable, and decentralization.

3.1 Objectives

O1. **Dataset exploration:** Explore existing Bluesky post datasets, with the option to construct new datasets if existing ones are insufficient, ensuring high-quality human and LLM-assisted annotations that capture disagreement and uncertainty.

O2. **Model evaluation:** Benchmark LLMs and transformer baselines in zero-shot, few-shot, and fine-tuned settings, including multilingual performance.

O3. **Tag grouping:** Develop LLM-enhanced clustering and graph-based tag grouping methods that account for cross-instance and cross-feed variation.

O4. **Bias and fairness:** Measure and mitigate demographic and political biases in LLM sentiment predictions.

3.2 Research Questions

- **RQ1:** How well do LLMs generalize from centralized datasets to Bluesky in terms of accuracy, calibration, and robustness to platform-specific language and tags?
- **RQ2:** How can tag groupings be modeled with LLM embeddings and graphs, and how stable are they across instances and feeds?
- **RQ3:** What social or demographic biases appear in LLM sentiment predictions, and how effectively can mitigation techniques reduce them?

4 Proposed Methodology

4.1 Data Collection and Preprocessing

Data collection from the Bluesky firehose will be used if existing datasets are insufficient. Stratified sampling will cover time periods (e.g., major events), topics (science, politics, health, entertainment), languages (English, Japanese, and others with sufficient volume), and observable feed generators. Preprocessing will include tokenization, emoji handling, language identification, and bot/spam detection (Sahneh et al., 2025). BlueSky firehose is generally available for retrieving data, except for private accounts where access is restricted due to user permissions and privacy settings.

4.2 Dataset Construction and Annotation

We will explore existing datasets, such as the Bluesky Social Dataset and POLITISKY24, for sentiment and hashtag analysis. If these datasets are insufficient, we will construct our own dataset from the Bluesky firehose. Two primary datasets will be developed: (i) a text-only sentiment dataset with post-level labels (positive, negative, neutral, and optionally finer-grained categories such as concern or trust), and (ii) a multimodal dataset where labels reflect combined text-image interpretation (Jin et al., 2024). Annotation will follow clear, platform-specific guidelines (Kadriu et al., 2022), use redundant labeling to capture ambiguity, employ LLM-assisted pre-labeling, and record annotator confidence for uncertainty-aware labels (Xiao et al., 2023).

4.3 Modeling

Sentiment models. We will evaluate general-purpose LLMs (zero-/few-shot with prompt engineering) and transformer baselines (BERT, BERTweet, domain-specific variants), using macro-F1, calibration, and robustness checks.

Tag groupings. We will build co-occurrence graphs of tags and posts, compute embeddings via LLM encoders, cluster tags with graph/community detection, and generate human-readable labels using LLMs. Temporal evolution of tag clusters and sentiment distributions will be tracked (Djenouri et al., 2019; Liou et al., 2020; Feng et al., 2015).

Multimodal modeling. We will apply multimodal LLMs and image–text fusion for sentiment classification and explanation, including late fusion and uncertainty-aware calibration. We will evaluate performance on posts dominated by visual versus textual sentiment (Jin et al., 2024; Pan et al., 2024).

4.4 Bias, Fairness and Uncertainty

Bias will be probed using paired prompts that vary demographic, political, or identity markers, with metrics capturing group differences and calibration (Venugopal et al., 2023). Mitigation techniques include adversarial training, representation debiasing, and post-hoc calibration (Corizzo et al., 2024). Uncertainty methods such as temperature scaling, Monte Carlo dropou,t and ensembles will flag unreliable predictions (Xiao et al., 2023).

5 Work Plan & Expected Contributions

The project will advance LLM-based sentiment analysis for Bluesky through: (i) analyze the AT Protocol and establish an ethical data collection pipeline; (ii) create and annotate text and multimodal datasets with uncertainty metadata; (iii) implement and benchmark LLM and transformer models, including tag grouping and multimodal approaches; (iv) evaluate bias, fairness and uncertainty and develop mitigation strategies (Huang, 2024); and (v) simulate sentiment- and tag-driven feed algorithms to study social effects. Key deliverables include anonymized Bluesky datasets with annotation guidelines, benchmark results and code, algorithms for tag groupings and their dynamics, bias and fairness analyses, and a simulation framework for feed evaluation. All work will follow privacy rules and ethical guidelines, keeping things clear and respecting user control. Contributions include providing the first Bluesky-specific sentiment benchmark and methods for LLM-enhanced tagging in decentralized environments.

References

W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. *Findings of the Association for Computational Linguistics: NAACL*. https://doi.org/10.48550/arXiv.2305.15005

L. He, S. Omranian, S. McRoy, and K. Zheng. 2024. Using Large Language Models for Sentiment Analysis of Health-Related Social Media Data: Empirical Evaluation and Practical Tips. *medRxiv* preprint. https://doi.org/10.1101/2024.03.19.24304544

M. Nasution et al. 2023. Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. University of Islam Riau repository. https://doi.org/10.1109/ACCESS.2025.3574629

X. Fu et al. 2023. Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study *PubMed*-indexed journal. https://doi.org/10.2196/51069

M. Kleppmann et al. 2024. Bluesky and the AT Protocol: Usable Decentralized Social Media. *ACM* https://doi.org/10.1145/3694809.3700740

E. Sahneh, G. Nogara, M. DeVerna, N. Liu, L. Luceri, F. Menczer, F. Pierri, and S. Giordano. 2025. The Dawn of Decentralized Social Media: An Exploration of Bluesky's Public Opening. ISBN: 978-3-031-78540-5. pp.422-437 https://doi.org/10.1007/978-3-031-78541-2_26.

T. Huang. 2024. Decentralized social networks and the future of free speech online. *Computer Law & Security Review*, 55:106059. https://doi.org/10.1016/j.clsr.2024.106059.

Y. Djenouri, A. Belhadi, and J. C. W. Lin. 2019. Deep learning based hashtag recommendation system for multimedia data *Information Processing & Management*. https://doi.org/10.1016/j.ins.2022.07.132

H.-T. Liou et al. 2020. Dynamic Graph Transformer for Implicit Tag Recognition. In *Proceedings of ACL*. https://doi.org/10.18653/v1/2021.eacl-main.122

W. Feng et al. 2015. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. *Proceedings - International Conference on Data Engineering*, 2015, 1561-1572. https://doi.org/10.1109/ICDE.2015.7113425.

X. Jin et al. 2024. MM-Soc: A Comprehensive Benchmark for Multimodal LLMs on Social Media. In *Proceedings of ACL*. https://doi.org/10.48550/arXiv.2402.14154

Q. Pan and Z. Meng. 2024. Hybrid Uncertainty Calibration for Multimodal Sentiment Analysis. *Electronics*. https://doi.org/10.3390/electronics13030662.

T. Xiao et al. 2022. Uncertainty Quantification and Calibration for Pre-Trained Language Models. In *Findings of ACL*. https://doi.org/10.48550/arXiv.2210.04714

M. I. Radaideh, O. H. Kwon, and M. I. Radaideh. 2025. Fairness and Social Bias Quantification in Large Language Models for Sentiment Analysis. *Knowledge-Based Systems*, 319:113569. https://doi.org/10.1016/j.knosys.2025.113569.

J. P. Venugopal, A. A. Subramanian, G. Sundaram, M. Rivera, and P. Wheeler. 2023. A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data. *Applied Sciences*, 14(23):11471. https://doi.org/10.3390/app142311471.

S. Vallejo Vera and H. Driggers. 2025. LLMs as Annotators: The Effect of Party Cues on Labelling Decisions by Large Language Models. *Humanities and Social Sciences Communications*, 12:1530. https://doi.org/10.1057/s41599-025-05834-4.

R. Corizzo and F. S. Hafner. 2024. Mitigating Social Bias in Sentiment Classification via Ethnicity-Aware Algorithmic Design. *Social Network Analysis and Mining*, 14:208. https://doi.org/10.1007/s13278-024-01369-9.

A. Kadriu et al. 2022. Human-annotated dataset for social media sentiment analysis for Albanian language. *Diva Portal* technical report. https://doi.org/10.1016/j.dib.2022.108436

P. Rostami, V. Rahimzadeh, A. Adibi, and A. Shakery. 2025. POLITISKY24: U.S. Political Bluesky Dataset with User Stance Labels [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15616911.

A. Failla and G. Rossetti. 2025. Bluesky Social Dataset [Data set]. Zenodo. `https://doi.org/10.5281/zenodo.14669616`.

F. N. Silva, K.-C. Yang, W. Zhao, and B. Tran Truong. 2024. Data for: Exploring Emerging Social Media: Acquiring, Processing, and Visualizing Data with Python and OSoMe Web Tools [Data set]. Zenodo. https://doi.org/10.5281/zenodo.12748042.