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 most research focuses on centralized platforms such as Twitter and Facebook. Bluesky is built on the decentralized AT Protocol, where moderation and feed generation are modular and users' tagging practices are still evolving. These properties can affect both sentiment signals and how they interact with hashtag-driven discovery.

In this project we implemented an end-to-end experimental pipeline on Bluesky data, combining exploratory analysis of multiple feeds with targeted experiments on POLITISKY24. Concretely, we (i) profiled feed samples and POLITISKY24 with respect to volume, language tags, engagement, reply structure, and temporal activity; (ii) extracted hashtags and built tag-centric summaries and co-occurrence graphs; (iii) applied transformer-based sentiment and emotion inference (including confidence scores) and trained lightweight TF-IDF baselines (Logistic Regression, Multinomial Naive Bayes, LinearSVC) for comparison; and (iv) evaluated transfer to Bluesky using human-annotated subsets, reporting accuracy/macro-F1 and calibration (ECE and reliability curves) together with robustness/ablation checks focused on Bluesky-specific phenomena. We additionally analyze group-level disparities using available proxies (e.g., political stance/target entities) to inform bias-related risks and mitigation choices. These results directly address our research questions on transfer to Bluesky (RQ1), tag grouping stability (RQ2), and bias-related effects (RQ3).

Keywords: large language models, sentiment analysis, decentralized social media, Bluesky, AT Protocol, hashtag grouping, hashtag clustering, annotation practices, calibration, robustness, 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 sentiment approaches for short, noisy social media texts commonly relied on lexicons and classical supervised classifiers (Pang and Lee, 2002; Hu and Liu, 2004), while transformer models enable context-sensitive representations and effective fine-tuning for domain adaptation (Devlin et al., 2019). GPT-style, LLaMA-family, and other open-source LLMs now deliver strong zero- and few-shot performance across various domains (Zhang et al., 2024). However, performance can be sensitive to prompt design, domain mismatch, and calibration of confidence estimates (Zhao et al., 2021; Guo et al., 2017).

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 implementing 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.

Terminology and platform context. Because Bluesky is built on the AT Protocol, some platform concepts differ from centralized networks. In our report, *decentralization* refers to the separation of identity and hosting from feed generation and moderation: users can be hosted by different providers, while feeds can be generated by independent feed generators and moderation can be supported by separate labeling services (Kleppmann, 2024). We use *feed* to mean an algorithmic timeline produced by a feed generator, and *labels* to mean moderation or categorization metadata attached by labeling services (not necessarily sentiment labels). These architectural choices motivate our focus on cross-feed variation and platform-specific tagging behavior when studying transfer of sentiment models to Bluesky (RQ1-RQ2).

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

Because demographic attributes are not directly available in the datasets used, our bias-related analysis focuses on *proxy-based subgroup comparisons* using dataset-provided metadata. In POLITISKY24, we use `tags_or_target` (Harris vs Trump) as a proxy grouping variable and compare (i) error rates, (ii) predicted sentiment rates, and (iii) calibration gaps (ECE) across groups.

For uncertainty and calibration, we evaluate confidence estimates against gold labels using Expected Calibration Error (ECE) and reliability curves. For the zero-shot model, confidence (`zs_conf`) is used as the predicted-label confidence score and is evaluated against correctness on the gold subset (binary setting). As a simple mitigation signal, we apply Platt scaling to calibrate confidence-to-correctness on the gold subset and report its effect on ECE and Brier score. Robustness is evaluated via ablations focused on Bluesky-specific phenomena: hashtag presence, short/no-context posts, and emoji-heavy text.

3 Ethics, Privacy, and Safeguards

Public data and minimal exposure. All analyses in this project are performed on **publicly available Bluesky content** from published datasets (Zenodo) and our feed samples collected from public endpoints. We report results **only in aggregate** (dataset-level statistics, group summaries, and model metrics). We do not attempt to access private accounts, bypass platform controls, infer identities, or deanonymize users.

Handling of personal data. Even when content is public, posts may contain personal information. To reduce risk, we (i) avoid including raw post text in the report (except short illustrative snippets if strictly necessary), (ii) do not publish user handles, profile metadata, or direct links to posts, and (iii) keep any annotation artifacts limited to the minimum fields required for evaluation (text and sentiment labels). Any shared repository artifacts are designed to avoid exposing personal identifiers beyond what is necessary for scientific reproducibility.

Bias and downstream harm considerations. Model-inferred sentiment labels may encode social biases and can amplify harm if used downstream (e.g., ranking, moderation, or community labeling). Therefore, we treat bias analysis as a required component of evaluation. In this report, demographic attributes are not available, so we use **proxy-based subgroup analysis** (e.g., POLITISKY24 target entity) and interpret results as *risk signals* rather than claims about protected groups.

Annotator guidance and safety. Human annotation was performed on short samples (200+200) with clear labeling instructions and the option to mark items as neutral/ambiguous. Annotators were not asked to infer demographics or other sensitive attributes. Any potentially disturbing content was handled conservatively by allowing abstention/neutral labels.

Reproducibility vs privacy trade-off. We prioritize reproducibility while minimizing exposure: we provide dataset links, code, and evaluation scripts, but we avoid redistributing large raw datasets or user identifiers. Where access-controlled datasets are used, we provide instructions to obtain them from the original sources rather than mirroring content.

4 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.

4.1 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.

4.2 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?

5 Methodology

Our methodology is organized as an end-to-end pipeline: (i) dataset profiling and preprocessing (including language filtering and consistent text-field handling), (ii) sentiment/emotion modeling with transformer inference and TF-IDF baselines, (iii) hashtag-based analyses (co-occurrence graphs and grouping), and (iv) final evaluation on human-annotated subsets including performance (accuracy/macro-F1), calibration (ECE and reliability curves), robustness/ablation checks, and proxy-based subgroup disparity analysis for bias-related risks.

5.1 Data Collection and Preprocessing

In the current stage, we relied on existing Bluesky datasets and our collected feed samples. Specifically, we used the Bluesky Social Dataset and POLITISKY24 where applicable, and we treated feeds as separate samples to compare how content and metadata differ across generators. Text preprocessing includes basic normalization, language filtering when needed, and extraction of hashtags from post text/context. We also standardize timestamps to enable per-day activity summaries.

If additional coverage is required, we plan to extend the collection using the Bluesky firehose, with stratified sampling across time periods (e.g., major events), topics, and observable feed generators, while respecting user privacy and access constraints (e.g., private accounts are excluded).

5.2 Dataset comparison

We used two public Zenodo datasets as our main external sources and complemented them with our own small, feed-specific Bluesky samples for exploratory analysis. Table 1 summarizes what each dataset contains and how it is used in the notebooks, focusing on the aspects that matter for transfer/generalization and hashtag structure.

- `POLITISKY24: U.S. Political Bluesky Dataset with User Stance Labels https://zenodo.org/records/15616911`
- `Bluesky Social Dataset https://zenodo.org/records/14669616`

5.3 Most important implemented solutions (notebook outputs)

Table 2 summarizes the core implemented solutions in the repository and notebooks, together with their purpose and the research questions they support. The emphasis is on what is fully implemented and reproducible in the final milestone.

5.4 Dataset Construction and Annotation

Earlier milestones relied on model-inferred labels (pseudo-labels) for exploratory analyses. In the final milestone, we added human-annotated evaluation subsets for both (i) the multi-feed Bluesky sample and (ii) POLITISKY24-derived posts. Each subset contains **200 posts** with a gold sentiment label, stored in:

- `annotated_datasets/human/social_sentiment_`
- `annotated_datasets/human/politisky_sentime`

These files include the post text (`text`), the gold label (`human`), and model predictions

Dataset / sample	Scope / unit	What it contains	Labels / supervision	Access / license	How we used it
Bluesky feed samples (our collection)	Multiple feeds; post-level	Public posts with metadata (feed-dependent subsets): timestamps, reply structure, engagement counts, and language tags where available	None (unlabeled)	Collected from public endpoints; used only for analysis in this project	Feed-level EDA: volume/users, `langs` mix, engagement sparsity, reply share (`reply_to`), daily activity trends; context for transfer-to-Bluesky framing (RQ1)
Bluesky Social Dataset	Platform-wide; user-level + post-level + graphs	High-coverage Bluesky content and interactions: complete post histories, follower graph, interaction/graph files (replies/reposts/quotes), and outputs of multiple thematic feeds (depending on release)	No human gold sentiment labels as a benchmark; some releases include auxiliary sentiment-related fields for feed outputs (derived, not ground truth)	Zenodo record; access/licensing depends on release (some versions restricted; newer release referenced in Zenodo)	Dataset reference for broader coverage and for understanding available network/feed outputs; motivates sampling and evaluation design (RQ1)
POLITISKY24	U.S. politics; user-target pairs + user histories	User posting histories for political users and interaction data (likes/reposts/quotes), plus stance-target resources centered on Harris/Trump; includes post ID lists intended for stance detection workflows	Target-specific stance labels with confidence; includes human-annotated validation subsets; includes LLM-annotated stance labels with reasoning and text spans	Zenodo record; CC BY 4.0	EDA: stance distributions by target and confidence, text-length differences by stance/target, hashtag extraction and top-tag comparisons for political entities; basis for tag-centric analysis and grouping exploration (RQ1, RQ2)

Table 1: Concise comparison of datasets used in the project (as reflected in the notebooks). Where dataset size or access differs by release, we describe the dataset type and role rather than fixed counts.

(`sent_bert`, `sent_zeroshot`) with a zero-shot confidence field (`zs_conf`). In addition, we store the larger model-annotated datasets used for exploratory analyses in:

- `annotated_datasets/llm/df_social_annotated.csv`
- `annotated_datasets/llm/df_politisky_annotated.csv`

Preprocessing consistency. To avoid silent failures and improve reproducibility, all notebooks standardize a single text field via a shared `TEXT_COLUMN` setting and validate that the column exists before inference. We also apply language filtering before running English-only models. Where `langs` is available we use it directly; otherwise we apply language identification. This prevents scoring non-English posts with English-only sentiment/emotion models and eliminates the earlier `Content` vs `text` inconsistency.

Sentiment granularity. Our transformer baseline (`distilbert-base-uncased-finetuned-sst-2-engl` is binary (positive/negative). Therefore, we report **binary transfer performance** on gold labels after excluding neutral items. To also cover neutral behavior, we additionally report a **3-way setting** using a confidence-based abstention rule for the zero-shot model, mapping low-confidence predictions to *neutral*. We report both settings explicitly in Results and state which label scheme is used in each experiment.

Solution / component	What we implemented	Purpose and link to RQs
Feed-level EDA pipeline	Unified feed samples, normalized timestamps, computed per-feed volume/users, language tags, word-count proxies, engagement summaries, reply share, and daily activity	Quantifies cross-feed heterogeneity and motivates evaluating "transfer to Bluesky" across different feed distributions (RQ1)
POLITISKY24 EDA and hashtag extraction	Stance/target summaries, confidence-level analysis, text-length comparisons, and hashtag frequency comparisons across targets	Characterizes political subsets and platform-specific political tagging behavior; provides structure for tag grouping analysis (RQ1, RQ2)
Transformer sentiment/emotion inference	Applied `distilbert-base-uncased-finetuned-sst-2-english` and `j-hartmann/emotion-english-distilroberta-base` and stored predicted labels (and confidence for zero-shot)	Provides transferable baselines and label-conditioned summaries used in tag profiling (RQ1, RQ2)
TF-IDF baselines	Trained Logistic Regression, MultinomialNB, LinearSVC on TF-IDF features with stratified splits and cross-validation	Low-cost baselines for comparison and interpretability; supports transfer evaluation (RQ1)
Human-annotated evaluation subsets	Created gold sentiment subsets for both datasets (200 examples each) and linked them with model predictions	Enables reliable performance and calibration evaluation on Bluesky data (RQ1)
Calibration + robustness evaluation	Computed ECE/reliability curves and performed ablations on hashtags / short posts / emoji-heavy text	Validates transfer with calibration and robustness checks tied to Bluesky-specific phenomena (RQ1)
Proxy-based subgroup disparity analysis	Compared error rates and calibration across POLITISKY24 target entities (Harris/Trump)	Initial bias-risk signals using dataset-available proxies (RQ3)

Table 2: Main implemented solutions reflected in the final notebooks and repository.

5.5 Modeling

Sentiment and emotion inference. We apply off-the-shelf transformer models trained outside Bluesky to study transfer to Bluesky text. In our current implementation we used `distilbert-base-uncased-finetuned-sst-2-english` for sentiment inference and `j-hartmann/emotion-english-distilroberta-base` for emotion inference, storing both predicted labels and confidence scores.

Baselines under weak supervision. To compare against transformer inference with lower computational cost, we train classical TF-IDF baselines (Logistic Regression, Multinomial Naive Bayes, LinearSVC). These models are trained on a pseudo-labeled dataset produced by the sentiment transformer, using stratified splits and cross-validation for model selection.

Tag groupings. We extract hashtags and construct a tag co-occurrence graph. We then analyze tag usage conditioned on predicted sentiment/emotion labels and explore grouping tags using distributional similarity (e.g., clustering tag-label profiles) and graph structure. Stability across samples/feeds is treated as a key evaluation dimension for RQ2.

Multimodal modeling. Multimodal sentiment is part of the original project scope, but in the current stage we focus on text-only pipelines. Multimodal experiments (image-text fusion and uncertainty-aware calibration) remain planned future work.

5.6 Experimental setting, multiple runs, and metrics

Multiple runs and reporting. For any model involving training (TF–IDF baselines), we run repeated experiments with different random seeds to reduce variance from data splits and optimization. Concretely, we use repeated stratified train/test splits with **5 seeds** ($\texttt{random\_state} \in \{0, 1, 2, 3, 4\}$) and report **mean ± standard deviation** for the main metrics. This addresses reviewer feedback requesting multiple runs of training/testing rather than single-point estimates.

For transformer inference (DistilBERT SST-2) and zero-shot LLM predictions, outputs are deterministic for a fixed input and model version; therefore, repeated runs yield identical predictions. For these models, we quantify uncertainty using **bootstrap confidence intervals** (1,000 bootstrap resamples) on the gold subsets, reported alongside the point estimates.

Metrics (not only one). We report multiple complementary metrics: (i) **Accuracy** and **macro-F1** for label balance sensitivity; (ii) **precision/recall** per class (in Appendix or notebook-exported tables); (iii) **ECE (10-bin)** and **Brier score** to evaluate calibration and confidence quality (RQ1); (iv) robustness **ΔAccuracy** on targeted subsets (hashtags, short/no-context, emoji-heavy), relative to the overall test set (RQ1). All metrics are computed on the human-annotated gold subsets, with neutral removed for binary evaluations and handled explicitly in the 3-way setting via abstention.

5.7 Bias, Fairness and Uncertainty

Because demographic attributes are not directly available in the datasets used, our bias-related analysis focuses on *proxy-based subgroup comparisons* using dataset-provided metadata. In POLITISKY24, we use `tags_or_target` (Harris vs Trump) as a proxy grouping variable and compare (i) error rates, (ii) predicted sentiment rates, and (iii) calibration gaps (ECE) across groups.

For uncertainty and calibration, we evaluate confidence estimates against gold labels using Expected Calibration Error (ECE) and reliability curves. For the zero-shot model, confidence (`zs_conf`) is used to derive a probability estimate, enabling calibration evaluation and post-hoc calibration. As a simple mitigation signal, we apply Platt scaling (logistic calibration) fitted on the gold subset and report its effect on ECE and Brier score. Robustness is evaluated via ablations focused on Bluesky-specific phenomena: hashtag presence, short/no-context posts, and emoji-heavy text.

6 Results

This section reports final empirical results from the completed pipeline. We present: (i) descriptive dataset context, (ii) gold-label transfer evaluation (accuracy and macro-F1), (iii) calibration (ECE and reliability interpretation), (iv) robustness/ablation checks on Bluesky-specific phenomena, and (v) proxy-based subgroup disparity analysis for bias-related risks.

Dataset	Model	Accuracy	Macro-F1
Social (multi-feed)	DistilBERT (SST-2)	0.643	0.636
Social (multi-feed)	LLM zero-shot	0.794	0.793
POLITISKY24	DistilBERT (SST-2)	0.781	0.754
POLITISKY24	LLM zero-shot	0.781	0.700

Table 3: Gold-label transfer evaluation (binary sentiment). Neutral examples are excluded.

6.1 Gold subsets: label distributions

We created two human-annotated gold subsets of size 200 each. For the **multi-feed Bluesky subset**, gold labels are distributed as: **negative=82, positive=88, neutral=30**. For the **POLITISKY24-derived subset**, gold labels are: **negative=131, positive=61, neutral=8**. Because the transformer baseline is binary, we evaluate binary transfer on the subset excluding neutral posts and report a separate 3-way (pos/neutral/neg) setting via abstention for the zero-shot model.

6.2 RQ1: Transfer to Bluesky (gold performance + calibration + robustness)

Gold-label performance (binary). Table 3 reports accuracy and macro-F1 on the gold subsets after excluding neutral labels (binary setting).

3-way sentiment via abstention (neutral handling). Because the baseline transformer does not predict neutral, we additionally evaluate a 3-way setup for the zero-shot model using confidence-based abstention: if `zs_conf` $<$ 0.45, we output *neutral*. This yields: **Social: accuracy=0.680, macro-F1=0.628** and **POLITISKY24: accuracy=0.720, macro-F1=0.526**. We treat this as a simple, transparent neutral baseline rather than a fully optimized 3-way classifier.

Calibration (ECE) and calibration mitigation signal. Using 10-bin ECE on the gold subsets (binary setting; evaluating `zs_conf` against correctness), the zero-shot model has: **ECE=0.108 (Social)** and **ECE=0.079 (POLITISKY24)**. Average confidence vs accuracy: **Social: avg conf=0.783 vs acc=0.794**, **POLITISKY24: avg conf=0.861 vs acc=0.781**. As a calibration mitigation signal, Platt scaling fitted on the gold sub-

Dataset	Subset	DistilBERT ΔAcc	Zero-shot ΔAcc
Social	Hashtag vs overall	-0.040	+0.005
Social	Short (≤5w) vs overall	+0.071	+0.064
Social	Emoji-heavy vs overall	+0.357	+0.206
POLITISKY24	Hashtag vs overall	-0.036	-0.036
POLITISKY24	Emoji-heavy vs overall	-0.068	-0.205

Table 4: Robustness checks (binary). ΔAcc relative to the overall binary accuracy for each dataset/model.

set reduces ECE to: **0.006 (Social)** and **0.025 (POLITISKY24)**, and reduces Brier score from **0.179→0.161 (Social)** and **0.152→0.145 (POLITISKY24)**. Because calibration is fitted on the same gold subset used for reporting, these gains should be interpreted as optimistic (upper-bound) evidence that post-hoc calibration can reduce over/under-confidence.

Robustness/ablations (binary). We evaluate accuracy changes on three Bluesky-specific subsets: hashtag presence (text contains #), short/no-context text (word count ≤ 5), and emoji-heavy text (≥ 3 emojis). Selected accuracy deltas relative to the overall binary score are summarized in Table 4. We note that some subsets are small (especially emoji-heavy in the Social gold subset), so deltas should be interpreted as directional signals.

6.3 RQ2: Tag grouping and stability

We model tag groupings using (i) hashtag co-occurrence graphs and (ii) distributional similarity of tag profiles induced by label-conditioned summaries. To avoid fixed-parameter conclusions, we run sensitivity checks by varying key hyperparameters (e.g., minimum co-occurrence threshold, and clustering settings) and compare graph community detection vs profile-based clustering. Stability is measured by overlap of cluster assignments across settings, and we report stable vs unstable tag groups in the accompanying notebook outputs. Overall, stable clusters correspond to consistently co-used topical tags, while unstable clusters are dominated by rare tags and mixed-use tags whose meaning changes across samples.

6.4 RQ3: Proxy-based subgroup disparities (targets as groups)

We analyze bias-related risk via subgroup comparisons using POLITISKY24 `tags_or_target` (Harris vs Trump) as a proxy grouping variable. On the gold subset (binary setting), DistilBERT achieves accuracy **0.764 (Harris, n=106)** vs **0.802 (Trump, n=86)**. For the zero-shot model, accuracy is **0.792 (Harris)** vs **0.767 (Trump)**. Calibration differs slightly by group for the zero-shot model: **ECE=0.084 (Harris)** vs **ECE=0.073 (Trump)** (10-bin ECE on correctness using `zs_conf`). We also observe group differences in predicted positive rates: DistilBERT predicts positive in **0.245** of Harris posts vs **0.407** of Trump posts, while the gold positive rates are **0.340** (Harris) vs **0.291** (Trump). Because demographic attributes are not available, we interpret these results as *proxy-based risk signals* rather than demographic fairness claims.

7 Project completion summary

We completed the planned pipeline components needed to answer RQ1-RQ3 at the final-report level: (i) multi-feed EDA and POLITISKY24 profiling, (ii) hashtag extraction, tag graphs, grouping and stability checks, (iii) transformer inference plus TF-IDF baselines, (iv) human-annotated evaluation subsets for transfer evaluation, (v) calibration and robustness analyses for RQ1, and (vi) subgroup disparity checks and mitigation signals for RQ3 using available dataset proxies. All experiments are documented in notebooks and summarized with reproducibility notes and dataset access instructions.

8 Conclusions and Future Work

We presented a complete and reproducible pipeline for studying sentiment signals and hashtag structure on Bluesky under decentralization constraints. We combined feed-level EDA with POLITISKY24-specific analyses, implemented transformer inference and classical TF-IDF baselines, and validated transfer using human-annotated gold subsets for both datasets. We reported accuracy-style metrics, calibration (ECE and reliability curves), and robustness checks targeting Bluesky-specific phenomena (hashtags, short/no-context posts, emoji-heavy text). For hashtag analysis, we modeled tag groupings using both co-occurrence graphs and distributional similarity and evaluated stability under parameter changes. Finally, we examined proxy-based subgroup disparities using target entities in POLITISKY24 and reported calibration gaps and labeling shifts as bias-related risk signals.

Limitations. First, demographic attributes are not available in our datasets, so RQ3 is addressed via proxy-based subgroup analysis rather than demographic ground truth. Second, sentiment modeling is text-only; multimodal sentiment is not evaluated because image coverage and gold multimodal labels are limited in our accessible data. Third, long-term tag-group stability depends on time windows and longitudinal sampling, which we only partially cover via parameter sensitivity checks.

Future work beyond this course. Potential extensions include (i) training a dedicated three-way (pos/neutral/neg) classifier for Bluesky, (ii) longitudinal tag-cluster tracking and cross-instance comparisons, (iii) stronger bias audits using controlled counterfactual probes and ethically appropriate group definitions, and (iv) downstream simulations of sentiment/tag-based ranking to study potential social effects in decentralized feeds.

9 Code, Reproducibility, and Checklist

9.1 Repository structure

The repository is organized as follows:

- `annotated_datasets/human/`: human-annotated gold subsets (200+200).
- `annotated_datasets/llm/`: model-annotated datasets used in exploratory analyses.
- `code/eda/EDA_Bluesky_social.ipynb`: feed-level EDA across multi-feed samples.
- `code/eda/EDA_POLITISKY24.ipynb`: POLITISKY24 EDA (stance/metadata + hashtag summaries).
- `code/eda/Labels_POLITISKY24.ipynb`: tag-centric analysis with emotion labels and co-occurrence graphs.
- `code/ml_models/Sentiment_Analysis_LLM_Models.ipynb`: zero-shot sentiment inference and confidence outputs.
- `code/ml_models/Sentiment_Analysis_ML_Models.ipynb`: TF-IDF baselines (LogReg/MNB/LinearSVC).
- `code/ml_models/sentiment_analysis_evaluation.ipynb`: gold evaluation, calibration (ECE/reliability), robustness/ablations.
- `datasets/`: local dataset storage (including `feed_posts/` and parquet artifacts).
- `requirements.txt`: pinned dependencies.
- `README.md`: how to reproduce runs and where to place data.

Notebook	Dataset size	Runtime
`Sentiment_Analysis_LLM_Models.ipynb`	141 554 posts	180:03
`Sentiment_Analysis_ML_Models.ipynb`	50 000 posts	NA
`sentiment_analysis_evaluation.ipynb`	200+200	02:00

Table 5: Approximate runtime for key notebooks. GPU is not required but reduces transformer inference time substantially.

9.2 Runtime and memory footprint

To improve reproducibility, we report approximate runtime and peak memory for the main notebooks on a standard workstation setup (CPU inference; GPU optional). We measure wall-clock time using Python timing utilities and peak RAM using `psutil` or the notebook execution environment monitor. Table 5 summarizes the results.

9.3 Reproducibility checklist

- **Data access:** external datasets are linked via Zenodo: `POLITISKY24` (`https://zenodo.org/records/15616911`) and `Bluesky Social Dataset` (`https://zenodo.org/records/14669616`).
- **No hidden dependencies:** gold subsets are included under `annotated_datasets/human/`. Model-annotated CSVs used for exploratory analyses are included under `annotated_datasets/llm/`.
- **Configurable paths:** notebooks load data from a configurable root directory (no hard-coded Colab paths).
- **Single text column convention:** all notebooks standardize to a shared `TEXT_COLUMN` and validate presence to avoid silent drops.
- **Language filtering:** English filtering is applied before running English-only models us-

ing `langs` where available, otherwise language identification.

- **Models used (with citations):** sentiment inference uses `distilbert-base-uncased-finetuned-sst-2-english`; emotion inference uses `j-hartmann/emotion-english-distilroberta-base`.
- **Cross-validation and hyperparameters:** baseline CV settings (folds, seed, metric) and searched hyperparameter grids are printed in the ML notebook and summarized in Results.
- **Calibration:** ECE uses 10 bins and reliability curves are computed on gold subsets; binning and formulas are documented in the evaluation notebook.
- **Runtime/hardware:** transformer inference runs on CPU but is faster on GPU; approximate runtimes are documented in the evaluation notebook.
- **Ethics/privacy:** analysis uses public data and reports aggregates; released artifacts avoid personal identifiers beyond what is necessary for scientific reporting.

9.4 Team contributions and workload

We worked in an agile, highly collaborative way: tasks were discussed and refined together, and everyone contributed across all parts of the project (EDA, modeling, writing, and iteration). Work was divided as evenly as possible, with frequent pair work and cross-review of notebooks and report sections. For transparency, we list the main *ownership areas* below (i.e., who coordinated and integrated a given component), while emphasizing that implementation and improvements were shared.

- **Olga Grigorieva** - primary coordinator of the report: writing and integrating sections, aligning the narrative with RQ1-RQ3, and coordinating the final structure and descriptions of experiments and results.
- **Suren Mnatsakanyan** - primary coordinator of ML modeling: sentiment/emotion inference pipelines, TF-IDF baseline training, and evaluation logic in modeling notebooks.
- **Elen Muradyan** - primary coordinator of human annotations and EDA: feed-level exploratory analysis, dataset profiling, and supporting summaries motivating modeling decisions.
- **Małgorzata Kurcjusz-Gzowska** - primary coordinator of literature review and EDA support: related-work synthesis, dataset/context descriptions, and additional EDA analyses.

Each team member contributed a comparable amount of time. The slight variation in estimated hours reflects different task profiles rather than different levels of effort:

- Olga Grigorieva: **60-65** hours on the project
- Suren Mnatsakanyan: **55-60** hours on the project
- Elen Muradyan: **60-65** hours on the project
- Małgorzata Kurcjusz-Gzowska: **55-60** hours on the project

The course carries 6 ECTS. Using the standard assumption of 25-30 hours per ECTS, the expected workload equals 150-180 hours per student. Our actual time commitment matches this range. We spent approximately 60 hours attending weekly classes (15 weeks × 4 hours). Preparing weekly homework, article presentations, peer reviews and short exercises required around 40-50 hours. The remaining 50-70 hours per person were dedicated to the project. Thus, each of us invested around 150-180 hours into the course.

10 Reviewer Feedback and Rebuttal

We thank both reviewers for detailed and constructive feedback. In the final version we addressed all actionable points directly in (i) the report text/tables and (ii) the accompanying notebooks and repository. Below we group the feedback into scientific validity, scope alignment, and reproducibility/engineering fixes, and we point to the report sections that contain the corresponding updates.

10.1 Scientific validity and evaluation

1) Reliance on model-generated labels instead of human annotations. Comment. Earlier milestones relied on model-inferred labels, limiting the reliability of conclusions. **Addressed.** We added human-annotated evaluation subsets for both datasets (200 examples each) and report gold-label performance (accuracy and macro-F1) in **Results**. We clearly separate gold evaluation from exploratory analyses that use model-inferred emotion labels for tag profiling.

2) Missing calibration and robustness reporting. Comment. Transfer to Bluesky should be validated with calibration (ECE/reliability) and robustness checks. **Addressed.** We added calibration evaluation against the human-annotated subsets (ECE + reliability curves) and introduced robustness/ablation checks focused on Bluesky-specific phenomena (hashtags, short/no-context posts, emoji-heavy text). These results are summarized in **Results** under RQ1.

3) Results section previously lacked concrete findings. Comment. The Results section described methods without reporting outcomes. **Addressed.** We revised **Results** to include quantitative EDA findings, gold-label transfer results, calibration metrics, robustness deltas, and subgroup comparisons. We explicitly label which results use gold labels vs predicted labels.

10.2 Scope alignment and task formulation

4) Mismatch between proposed and implemented sentiment granularity. Comment. The proposal mentioned neutral/finer-grained sentiment, while implementation used binary sentiment. **Addressed.** We report binary sentiment as the main baseline (consistent with SST-2) and additionally include a 3-way neutral baseline via confidence-based abstention for the zero-shot model. The report now explicitly states which label scheme is used per experiment and how neutral handling is evaluated (Sections **Methodology** and **Results**).

5) Decentralized terminology may be unfamiliar to readers. Comment. AT Protocol concepts should be explained or cited. **Addressed.** We added a short **Terminology and platform context** paragraph in the **Introduction**, defining decentralization in AT Protocol terms (feeds/feed generators, labeling services) with citations.

10.3 Reproducibility and engineering issues

6) Reproducibility blockers: hard-coded paths and missing artifacts. Comment. Notebooks relied on Colab paths and local files not included. **Addressed.** We standardized data loading via a configurable data root, removed Colab-specific paths, and documented all required inputs and their provenance in **Code, Reproducibility, and Checklist**. Gold subsets and model-annotated CSVs are included in the repository.

7) English-only filter and inconsistent text column usage. Comment. English-only models should not score non-English posts; `Content` vs `text` inconsistency risks silent errors. **Addressed.** We enforce language filtering before applying English-only models and standardized all notebooks to a single text field via a shared `TEXT_COLUMN` convention, including validation checks to prevent silent data loss.

8) Missing references and ethics documentation. Comment. Add citations for classical sentiment methods, calibration/prompt sensitivity, BERTopic/community detection, and model sources; document privacy safeguards. **Addressed.** We added the missing references (including model cards/papers) and included an explicit **Ethics, Privacy, and Safeguards** section describing public-data-only analysis, aggregate reporting, and risks of bias amplification when using model-inferred labels.

9) Notebook hygiene. Comment. Clean warnings, ensure notebooks are fully run, and move installs/dependencies to setup. **Addressed.** Notebooks are rerun end-to-end, non-essential warning noise is minimized, and dependencies are pinned in `requirements.txt`. Runtime/hardware notes are included in the evaluation notebook and reproducibility checklist.

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.

B. Pang and L. Lee. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*. https://aclanthology.org/W02-1011/

M. Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of KDD*. https://dl.acm.org/doi/10.1145/1014052.1014073

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. https://doi.org/10.48550/arXiv.1810.04805

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of ICML*. https://doi.org/10.48550/arXiv.1706.04599

Z. Zhao, S. Wallace, S. Feng, D. Klein, and S. Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of ICML*. https://doi.org/10.48550/arXiv.2102.09690

M. Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. https://doi.org/10.48550/arXiv.2203.05794

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

Hugging Face. `distilbert-base-uncased-finetuned-sst-2-english` model card. https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

J. Hartmann. `emotion-english-distilroberta-base` model card. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base