Sentiment Analysis with Large Language Models on Bluesky

Tag Groupings and Decentralized Social Media

Olga Grigorieva Małgorzata Kurcjusz-Gzowska
Elen Muradyan Suren Mnatsakanyan

Natural Language Processing - Project Proposal

Outline

Motivation & Context

Objectives & Questions

State of the Art

Research Gap

Dataset

PoC

Motivation & Context

Bluesky as a new environment

- AT Protocol separates:
  - **Identity** (DIDs)
  - **Data hosting** (repositories)
  - **Feeds & labels** (modular services)
- Multiple ranking, moderation, and labeling services can coexist
- Sentiment models interact with:
  - feed generators
  - labeling and moderation services
  - different hosting providers/instances
- Raises questions **Who controls the sentiment labels? Are they fair? Do they reflect the community's norms?**

Role of tags and hashtags

- Tags structure:
  - structure content discovery
  - help topic formation
  - shape community identity
- Prior work looked at:
  - deep models for hashtag recommendation
  - dynamic adaptation to semantic drift
  - graph-based tag clustering
- LLMs can go futher:
  - embed tags and posts in a shared semantic space
  - generate human-readable cluster labels
  - explain tag meanings and relationships

Objectives & Questions

Project objectives

1. Explore **Explore existing Bluesky post datasets**
   - text-only and multimodal (where applicable)
   - with human and LLM-assisted annotation
2. Benchmark **LLMs and transformer baselines**
   - zero-shot, few-shot, and fine-tuned
   - multilingual settings
3. Develop **LLM-enhanced tag groupings**
   - clustering + graph-based methods
   - tailored to AT Protocol's decentralised structure
4. Develop **Test Classical ML Models**
   - Use read HF twitter-roberta-base-sentiment-latest to label data
   - Compute Embedding Vectors
   - Train classical models on top: Logistic Regression, Naive Bayes, XGBoost
   - Tune the parameters of models, find best subset of features
5. Analyze **bias, fairness, and uncertainty**
   - probe demographic and political biases
   - evaluate mitigation strategies

Key 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?

State of the Art

State of the Art

LLMs for sentiment on social media

- Transformer based models and large language models now dominate sentiment analysis on social media
- LLMs reach strong results in zero shot and few shot setups on many benchmarks.
- They still struggle with complex structured tasks such as aspect based sentiment and opinion role extraction.

State of the Art

Domain specific and multilingual findings

- In domain specific settings, LLMs can match or outperform fine tuned transformers with good prompting.
- Performance drops on noisy or highly specialised content and for low resource languages.
- Studies highlight issues with sarcasm, emojis, code switching and non standard language.

State of the Art

Decentralised social media and Bluesky

- Bluesky uses the AT Protocol, which separates identity, storage and feed generation
- Moderation, labeling and ranking can be implemented by independent services and custom feeds
- Existing studies describe growth, language distribution and toxicity but rarely use LLM based sentiment

State of the Art

Hashtag groupings

- Hashtags organise topics, discovery and community identity on platforms like Twitter and Instagram
- Existing work rarely considers decentralised architectures or instance-specific tag semantics

Research Gap

Gaps in current literature

- No benchmarks of **LLM sentiment** on decentralized social media
- Tag recommendation and clustering:
  - mostly designed for centralized platforms
  - rarely use LLMs in the core modeling loop
- Bias/fairness studies focus on Twitter, finance, or specific domains
- No work linking:
  - LLM sentiment
  - tag groupings

Dataset

Dataset & annotation

Datasets

- **POLITISKY24:** U.S. Political Bluesky Dataset with User Stance Labels
  - labels: Trump/Harris
- **Bluesky Social Dataset:** Data on the individual posts collected
  - includes followers, likes, interactions, but no labels
- **Data From Bluesky and Mastadon:** For Exploring Emerging Social Media
  - includes embeddings

Annotation protocol

- Clear guidelines, platform-specific examples
- Redundant labeling (≥ 2 annotators/post)
- Capture disagreement vs. annotator error
- LLM-assisted pre-labeling with human verification
- Record confidence / uncertainty

Sentiment modelling

- **Models**
  - General LLMs (open + proprietary where allowed) and Classical ML Methods over Embeddings
  - Transformer baselines (BERT, BERTweet, domain-specific variants)
- **Setups**
  - Zero-shot and few-shot with prompt engineering
  - Supervised fine-tuning on Bluesky data
- **Evaluation**
  - macro-F1, accuracy
  - calibration (ECE, reliability curves)
  - robustness: paraphrasing, noise, domain shift
  - multilingual performance

Tag groupings with LLMs and graphs

- Build **tag-post graphs**
  - nodes: tags (hashtags, labels), posts, possibly users
  - edges: co-occurrence, reply/quote relations
- Compute embeddings:
  - LLM sentence/tag embeddings
  - graph-based representations
- Cluster tags:
  - community detection / clustering
  - LLM-generated names + descriptions for clusters
- Track:
  - temporal evolution of tag clusters
  - sentiment distributions per cluster
  - cross-instance and cross-feed differences

PoC

Work plan

Phase	Main activities
1. Platform & data	AT Protocol analysis, data pipeline, ethics approval
2. Datasets	Sampling, annotation guidelines, human & LLM annotation
3. Modelling	Classical ML Methods, LLM/transformer benchmarking, tag grouping
4. Bias & uncertainty	Probing, mitigation, calibration studies
5. Dissemination	Papers, code, datasets, documentation

Proof of Concept

- Initial data exploration on given datasets
- LLM-based annotation before human verification
- Different models for sentiment analysis

```python
import pandas as pd

df_results = pd.DataFrame({
    'text': texts_list,
    'emotion': pred_labels,
    'confidence': pred_scores
})

df_results.head()
```

	text	emotion	confidence
0	Assuming Harris is the pick. Biden might have ...	fear	0.401503
1	Gonna fucking Battle Royale us into a Trump pr...	anger	0.851509
2	Kamala/Harris Kamala Harris is very good at th...	neutral	0.690806
3	Not to get too political on here, but I really...	fear	0.288336
4	To any and ALL MAGA calling for unity and the ...	fear	0.616217

```python
df_results.to_csv("bluesky_emotions.csv", index=False)
```

Proof of Concept

85]:

	penalty	C	mean_f1_macro	std_f1_macro
7	l2	10.00	0.736636	0.005141
3	l1	10.00	0.724574	0.004607
6	l2	1.00	0.684856	0.003110
2	l1	1.00	0.675014	0.003632
5	l2	0.10	0.529786	0.004712
1	l1	0.10	0.510757	0.004042
4	l2	0.01	0.398399	0.000642
0	l1	0.01	0.398080	0.000018

References

Selected 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 As- sociation for Computational Linguistics
- L. He, S. Omranian, S. McRoy, and K. Zheng. 2024. Using Large Language Models for Sentiment Analysis of Health-Related Social Media Data: Empiri- cal Evaluation and Practical Tips. medRxiv preprint.
- M. Nasution et al. 2023. Benchmarking Open-Source Large Language Models for Sen- timent and Emotion Classification in Indonesian Tweets. University of Islam Riau repository.
- T. Huang. 2024. Decentralized social net- works and the future of free speech online. Computer Law Security Review, 55:106059.
- E. Sahneh, G. Nogara, M. DeVerna, N. Liu, L. Luceri, F. Menczer, F. Pierri, and S. Gior- dano. 2025. The Dawn of Decentralized So- cial Media: An Exploration of Bluesky's Public Opening.
- Y. Djenouri, A. Belhadi, and J. C. W. Lin. 2019. Deep learning based hashtag recommendation system for multimedia data Information Processing Man- agement.
- J. P. Venugopal, A. A. Subramanian, G. Sun- daram, M. Rivera, and P. Wheeler. 2023. A Comprehensive Approach to Bias Mitiga- tion for Sentiment Analysis of Social Me- dia Data. Applied Sciences

Thank you!

Questions & discussion