POLITISKY24: U.S. Political Bluesky Dataset with User Stance Labels

Peyman Rostami[1], **Vahid Rahimzadeh**[1], **Ali Adibi**[1], **Azadeh Shakery**[1,2]
[1]University of Tehran, Iran
[2]Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
{pe.rostami,rahimzade,adibialii,shakery}@ut.ac.ir

Abstract

Stance detection identifies the viewpoint expressed in text toward a specific target, such as a political figure. While previous datasets have focused primarily on tweet-level stances from established platforms, user-level stance resources—especially on emerging platforms like Bluesky—remain scarce. User-level stance detection provides a more holistic view by considering a user's complete posting history rather than isolated posts. We present the first stance detection dataset for the 2024 U.S. presidential election, collected from Bluesky and centered on Kamala Harris and Donald Trump. The dataset comprises 16,044 user-target stance pairs enriched with engagement metadata, interaction graphs, and user posting histories. POLITISKY24 was created using a carefully evaluated pipeline combining advanced information retrieval and large language models, which generates stance labels with supporting rationales and text spans for transparency. The labeling approach achieves 81% accuracy with scalable LLMs. This resource addresses gaps in political stance analysis through its timeliness, open-data nature, and user-level perspective. The dataset is available at https://doi.org/10.5281/zenodo.15616911.

1 Introduction

Stance detection is the task of automatically determining the viewpoint expressed in a text toward a specific target, such as an entity, topic, or claim (Mohammad et al., 2016a). This task serves as a vital component of many NLP tasks, such as fake news detection, fact-checking, rumor verification and user profiling (Rahimzadeh et al., 2025). Furthermore, it supports applications like identifying public opinion, tracking sentiment toward politicians or products, and understanding user attitudes in social media conversations (Zhang et al., 2024a; Zarharan et al., 2024).

Research in stance detection spans multiple domains, each with its unique challenges and characteristics. In the financial domain, datasets like WT-WT (Conforti et al., 2020) have focused on mergers and acquisitions between major companies. Health-related datasets emerged during the COVID-19 pandemic, capturing positions toward public health figures and policies (Glandt et al., 2021). General-purpose datasets such as VAST (Allaway and McKeown, 2020) and EZ-STANCE (Zhao and Caragea, 2024) have addressed broader topics spanning education, entertainment, and environmental protection. However, the political domain, particularly U.S. politics, has attracted the most attention due to its societal impact and complexity. Numerous datasets have been developed focusing on political figures and events, with a significant emphasis on the 2016 and 2020 U.S. presidential elections (Kawintiranon and Singh, 2021a; Li et al., 2021a; Mohammad et al., 2017, 2016a). To the best of our knowledge, no datasets currently exist for the 2024 election.

Twitter (now X) has traditionally been the primary source for stance detection datasets, but recent API restrictions have severely limited access

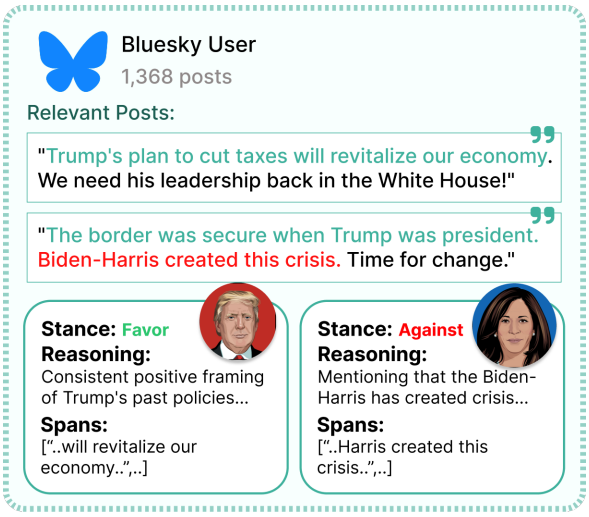

Figure 1: An illustrative example of the user stance with metadata in PolitiSky24.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21976–21993
November 4-9, 2025

to public data, hampering research efforts. Additionally, most existing datasets focus on individual post-level analysis, overlooking valuable contextual information from users' complete posting histories. Examining stance at the user-level provides a more comprehensive understanding of political viewpoints by considering broader expression patterns rather than isolated posts. Furthermore, while alternative platforms like Bluesky offer decentralized architectures and greater data accessibility, they remain largely unexplored for stance detection dataset creation. To address these limitations, we present a comprehensive stance dataset focused on the 2024 US presidential candidates, specifically targeting Kamala Harris and Donald Trump. Our contribution includes 16,044 user-target stance pairs with rich metadata and network information (See Figure 1). The dataset was created through a meticulously configured pipeline that first filters users with a hashtag-based algorithm to identify relevant accounts, then employs LLMs to determine stance based on high-quality chunks of user history. We developed carefully crafted validation datasets to evaluate different components, testing various embedding models for context retrieval and LLMs for stance labeling. Our thoroughly evaluated pipeline provides stance labels with reasoning explanations, relevant text spans, and complete network structures of users' interactions, namely repost, like and following. The dataset enables analysis of political discourse on emerging platforms while offering insights into LLM capabilities for understanding complex political communication.

The remainder of this paper is organized as follows: Section 2 reviews related work in both stance detection and studies of the Bluesky platform. Section 3 describes our Bluesky dataset, detailing the network structure and data collection methodologies that resulted in our US political dataset. Section 4 presents our stance labeling pipeline and the resulting stance dataset. Section 5 provides experimental results and comprehensive error analysis of LLM performance in stance detection and presents key insights derived from our dataset. Section 6 describes future outlooks in detail. Finally, Section 7 concludes the paper and discusses future research directions.

2 Related Work

Bluesky Social Network. Bluesky is a decentralized social network that has recently drawn research attention. Studies have examined its architecture (Kleppmann et al., 2024) and curated datasets to analyze user activity, network diversity, content moderation, and interaction patterns (Balduf et al., 2024; Jeong et al., 2024; Quelle and Bovet, 2025; Failla and Rossetti, 2024). Notably, analyses suggest that Bluesky currently has a predominantly left-leaning user base (Quelle and Bovet, 2025). However, it has not yet been explored for downstream tasks such as stance detection.

Stance Detection. Stance detection aims to identify a user's position toward a specific target and is widely studied across domains, such as politics (Li et al., 2021a), public health (Glandt et al., 2021), and multi-topic datasets (Mohammad et al., 2016a). For example, Mohammad et al. (2016a) introduced a dataset covering multiple topics, including U.S. political figures, and Conforti et al. (2020) investigated cross-domain stance detection, highlighting domain shift challenges. Recent efforts have also focused on zero-shot stance detection (ZSSD), where the target is unseen during model training. For example, EZ-STANCE (Zhao and Caragea, 2024) introduced a large-scale English ZSSD dataset containing 47,316 annotated text-target pairs, covering both claim- and noun-phrase-based targets across diverse domains.

Despite these advances, existing work has mainly focused on post-level (e.g., tweet) stance detection on centralized platforms, with limited attention to user-level stance detection on decentralized networks like Bluesky. Table 1 summarizes notable political stance datasets, most of which primarily focus on tweet-level annotations. Only a few include user-level annotations, and these rely on X/Twitter. Furthermore, to the best of our knowledge, there are currently no stance detection datasets centered on key figures or targets from the 2024 U.S. presidential election—a gap addressed by our proposed *PolitiSky24* dataset.

3 Bluesky Dataset on U.S. Politics

To create a dataset capturing Bluesky users' stances toward the main candidates in the 2024 U.S. presidential election, we first needed to identify a set of users engaged in U.S. politics, along with their posting histories. In this section, we describe the process of collecting this dataset and present statistics on users' posts and network connections. In the next section, we outline the steps taken to construct

Dataset	Targets	Labels	Source	Collection year(s)	Sample level	#HAS	#MAS
Li et al. (2021a)	Donald Trump, Joe Biden, Bernie Sanders	Favor, Against, Neutral	Twitter	2020	Tweet	21,574	0
Mohammad et al. (2016b)	Donald Trump, Hillary Clinton, Atheism, Climate change, Feminist movement, Abortion	Favor, Against, Neither	Twitter	2016	Tweet	4,870	0
Dutta et al. (2022)	US democratic party, US republican party, Bharatiya Janta Party, Indian National Congress, Aam Admi Party	Pro, Anti, Other	Twitter	2019–2020	Tweet	8,007	0
Kawintiranon and Singh (2021b)	Donald Trump, Joe Biden	Support, Oppose, Neutral	Twitter	2020	Tweet	2,500	0
Zhang et al. (2024b)	Donald Trump, Joe Biden	Pro	Twitter	2020	User	0	311,618
Darwish et al. (2020)	Donald Trump, Brett Kavanaugh, Recep Tayyip Erdoğan	Pro, Anti	Twitter	2018	User	150	296,556
PolitiSky24 (our dataset)	Donald Trump, Kamala Harris	Favor, Against, Neither	Bluesky	2024	User	445	8,467

Table 1: An overview of political stance detection datasets, detailing their targets, labels, sources, collection years, sample levels, and the number of human-annotated (#HAS) and machine-annotated (#MAS) samples.

the stance dataset based on this U.S. political data.

3.1 Bluesky Social Network

Bluesky is a decentralized social network where users can share short posts (up to 300 characters), along with images and videos, similar to Twitter/X (Balduf et al., 2024). On this platform, users can engage with others by following them, as well as by liking, reposting, quoting, and replying to their posts. A standout feature of Bluesky is its *Feed Generator* service, which enables users to create personalized timelines—called feeds—on various topics using custom algorithms (Jeong et al., 2024). This service provides real-time access to recent posts on Bluesky related to any topic of interest, such as U.S. politics.

3.2 Data Collection on U.S. Politics

The process of collecting data from the U.S. political domain on Bluesky is outlined as follows. First, three existing feeds focused on U.S. political content were identified (refer to Appendix D.1.1 for further information about these feeds). Posts from these feeds were then collected over a 16-day period, from November 12 to November 27, 2024. Next, individuals who had at least 10 distinct posts within the collected posts were defined as target users. For each target user, all posts they had ever published or reposted on Bluesky, along with the associated metadata, were collected. Finally, engagement data were gathered using the Bluesky API. This included information about users who liked, reposted, or quoted the target users' posts, as well as those who followed the target users or were followed by them.

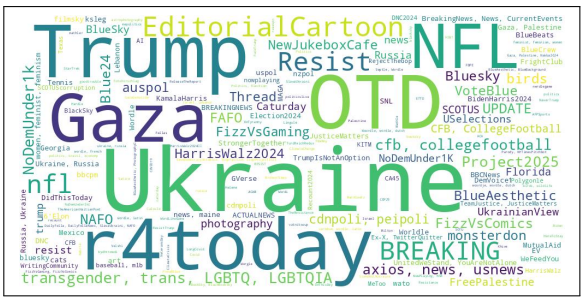

Figure 2: Hashtag cloud generated from users' posts.

3.3 General Statistics

3.3.1 Post statistics

Table 2 presents the overall statistics of the collected posts. In total, nearly 18.5 million posts were collected from 8,561 target users between November 2022 (the launch of Bluesky) and November 2024. Among these posts, 69.5% are original, while 30.5% are reposts. The dataset includes 226,273 unique hashtags. As shown in the hashtag cloud in Figure 2, these hashtags cover a wide range of topics, predominantly related to U.S. politics. This highlights the effectiveness of the data collection process in capturing discussions centered on U.S. political issues.

3.3.2 User Statistics and Network Connections

Drawing from our monitored political feeds discussed earlier, we identified 8,561 unique users

# Posts			Average post length		# Unique hashtags	Collection period
Total Posts	Original posts	Reposts	Words	Characters		
18,416,787	12,804,910	5,611,877	21.2	128.8	226,273	Nov. 2022* - Nov. 2024

* Bluesky launch time

Table 2: General statistics of the collected posts.

who contributed at least 10 posts/reposts during the specified analysis period. For these core politically active users, we gathered two types of engagement networks using the Bluesky API - a network of likes and a network of reposts between these users, with their characteristics presented in Table 3. The threshold of 10 unique posts/reposts ensures we capture consistently active users rather than occasional participants, providing a robust representation of the core political discourse network on the platform. Our analysis reveals two highly connected interaction networks with distinct characteristics. The Likes network, consisting of 8,454 nodes and 869,367 edges, demonstrates denser interaction patterns compared to the Reposts network (8,193 nodes, 498,084 edges). Both networks show exceptional connectivity, with largest connected components (LCC) encompassing over 99.6% of all nodes, indicating nearly complete network connectivity.

The networks exhibit strong small-world properties, characterized by high clustering coefficients (0.319 for Likes, 0.296 for Reposts) that significantly exceed those of comparable random networks (0.024 and 0.015 respectively). Combined with short average path lengths (2.185 and 2.363) and substantial small-world coefficients (11.817 and 17.925), these metrics indicate efficient information diffusion pathways within the networks. The centrality distributions show notable skewness, particularly in betweenness centrality (skewness of 19.16 and 21.89) and PageRank (13.67 and 7.05), suggesting the presence of key influential nodes in both interaction networks.

Interaction patterns differ between likes and reposts, with the Likes network showing more intensive engagement (average degree 205.67) compared to the Reposts network (average degree 121.588). This disparity likely reflects the different cognitive and social costs associated with these interaction types, where likes represent a lower-threshold engagement mechanism compared to reposts. While both networks exhibit strong connectivity, these differences in user behavior lead to distinct structural roles for each interaction type.

Metric	**Likes Net.**	**Reposts Net.**
Centralities (Mean)		
Betweenness	0.000180	0.000225
Closeness	0.386255	0.318847
Eigenvector	0.004776	0.004629
PageRank	0.000118	0.000122
In-Degree	117.260	64.089
Out-Degree	117.260	64.089
Small World Metrics		
Clustering Coef.	0.319	0.296
Avg Path Length	2.185	2.363
Random Clust.	0.024	0.015
Random Path	1.982	2.143
SW Coefficient	11.817	17.925
Network Metrics		
Nodes	8,454	8,193
Edges	869,367	498,084
Density	0.024	0.015
Avg Degree	205.670	121.588
LCC Size (%)	99.858	99.609

Table 3: Network Analysis Metrics

These structural roles manifest as a trade-off between dense community bonding and efficient information bridging. The Likes network functions to build "bonding capital" (Fernandez and Nichols, 2002), using its higher density to create a cohesive community where members are, on average, closer to all others (mean closeness centrality 0.387 vs. 0.319). Conversely, the Reposts network excels at building "bridging capital", with a higher mean betweenness centrality (0.000225 vs. 0.000180) that highlights its importance in linking different parts of the network. This demonstrates a superior efficiency: the Reposts network achieves nearly identical connectivity (LCC > 99.6%) with 43% fewer edges and a higher SW Coefficient (17.925 vs. 11.817). This suggests that the strategic, selective amplification of reposting creates a leaner but

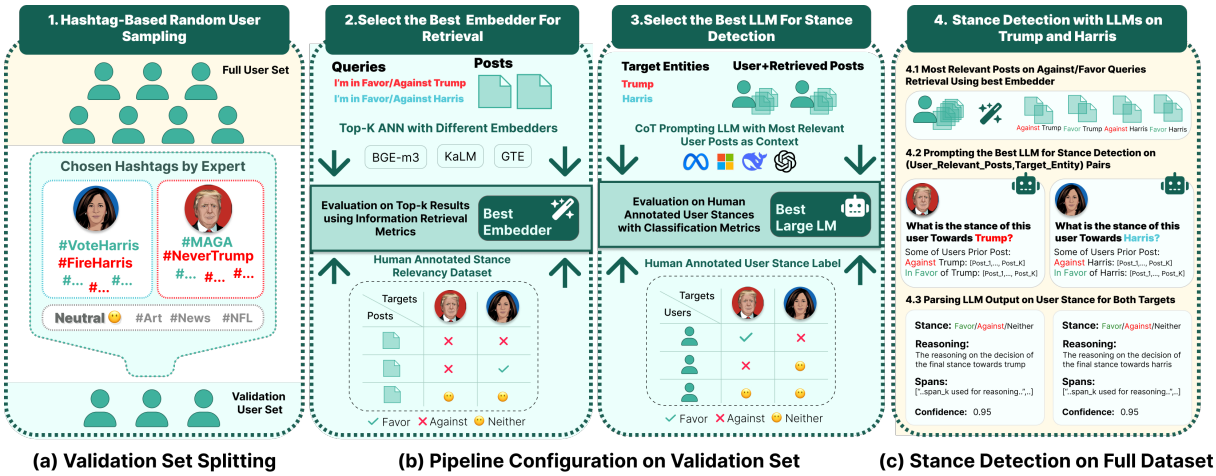

Figure 3: LLM-based annotation pipeline for labeling Bluesky users' stances toward Trump and Harris.

equally connected and more structurally efficient network for broad information dissemination.

For a more detailed analysis of the collected dataset, please refer to Appendix D.

4 Stance Detection

After collecting the Bluesky dataset related to U.S. politics in the previous section, we now focus on annotating the stances of the Bluesky users toward the key figures in the 2024 U.S. election. In this section, we detail the annotation process.

4.1 Task Definition

Our task is to create a labeled dataset for user-level, target-specific stance detection that identifies a user's stance—*Favor*, *Against*, or *Neither*—toward a specific target entity. In this study, the targets are Donald Trump and Kamala Harris, two key figures in the 2024 U.S. elections.

To construct the stance dataset, we retrieve up to 1,000 recent English-language posts for each user. These posts include original posts and reposts, excluding replies and quoted posts. Overall, the dataset encompasses 8,467 users[1] and 2,845,217 English-language posts.

4.2 Annotation Process

To label users' stances toward two target entities, Donald Trump and Kamala Harris, we begin by selecting approximately 5% of the users as the validation set through hashtag-based random sampling (Figure 3(a)). Next, we configure a two-step pipeline that uses this validation set to accurately label users' stances toward both targets using a LLM (Figure 3(b)). To mitigate hallucinations and reduce bias when annotating users' stances, the LLM should ground its answers in as much relevant data as possible. However, given that users often have extensive posting histories, we need a stance-based document retriever to generate concise and relevant contexts for the LLM. Therefore in the first step of our pipeline, we construct a human-annotated stance relevancy dataset from the validation set and evaluate several embedding models to identify the best-performing model for stance-based document retrieval. In the second step, we construct a human-annotated user stance dataset from the validation set and prompt various LLMs with the most relevant user posts as context to identify the most effective model for detecting user stances toward Trump and Harris. After configuring the pipeline using the validation set, we apply it to the full dataset to label each user's stance toward both targets, along with the reasoning behind the assigned label (Figure 3(c)).

The following subsections provide a detailed explanation of each part of the annotation process.

4.2.1 Validation Set Splitting

In this subsection, we describe the process used to select users for the validation set.

Following the approach of Li et al. (2021b), we used hashtags to identify validation users. Specifically, we consulted an expert to compile a list of hashtags drawn from users' posts across five categories: Favor-Trump, Against-Trump, Favor-Harris, Against-Harris, and Neither. The Neither category includes a broad range of hashtags re-

[1]Out of the initial 8,561 users, 94 were excluded due to the absence of any English-language content in their posts or reposts, resulting in a final set of 8,467 users.

Target	Against	Favor
Trump	#NeverTrump	#TrumpWillSaveAmerica
Harris	#FireKamala	#VoteHarrisWalz202

Example of Neither Hashtags: #BreakingNews, #Art, #Genocide, #2024election.

Table 4: Examples of hashtags for target-stance pairs.

lated to news, non-political topics, and political topics that do not convey a clear stance toward either Trump or Harris. Table 4 provides examples of the selected hashtags for each category. After finalizing the hashtag list, we used random sampling to select 445 users—approximately 5% of all users—who had used the identified hashtags in their posts, as the validation set.

4.2.2 Pipeline Configuration on Validation Set: Embedding Model Selection

In this subsection, we describe the process of constructing a human-annotated stance relevancy dataset from the validation set and using it to select the best embedding model for stance-based document retrieval.

Stance Relevancy Dataset Construction. Following the annotation guidelines provided in Table 5 in Appendix B, three annotators were tasked with selecting posts from the entire pool of validation users' posts in a distributed and balanced manner across all labels and target entities. This process yielded 350 post–target entity stance pairs. For further details, please refer to Appendix D.2.1.

After preparing the list of labeled documents, we constructed a validation stance relevancy dataset based on them. In this dataset, a document is assigned a relevance label of 1 for a given query if it expresses the stance specified in the query toward the target entity mentioned in that query; otherwise, it is assigned a label of 0. Each query corresponds to one of the following four options: 1) "I am in favor of Donald Trump," 2) "I am against Donald Trump," 3) "I am in favor of Kamala Harris," and 4) "I am against Kamala Harris." As a result, the stance relevancy dataset consists of a total of $4 \times 175 = 700$ query-stance relevance label pairs.

Embedding Model Selection. To identify the most effective embedding model for stance-based document retrieval, we evaluated several models on the stance relevancy dataset, which was specifically designed for this task. We used approximate nearest neighbor (ANN) search for document retrieval (Arya et al., 1998) and assessed model performance using standard information retrieval metrics. The model with the highest retrieval performance was selected to retrieve the most stance-relevant documents for each user–target entity pair.

4.2.3 Pipeline Configuration on Validation Set: LLM Selection

In this subsection, we describe how we constructed a human-annotated user stance dataset from the validation set and used it to identify the most effective LLM for user stance detection.

User Stance Dataset Construction. We asked three experts to evaluate each validation user's posts and label the user's stance toward each target entity. We achieved a Krippendorff's Alpha of 0.74 (Trump) and 0.79 (Harris), which represents a substantial level of agreement given the task's subjectivity. The annotation guidelines and the distribution of stance labels assigned by the experts for both target entities (Trump and Harris) are available in Appendices B and D.2.2, respectively.

LLM Selection. To identify the most effective LLM for user stance detection, we evaluated several models on our validation user stance dataset. Specifically, for each validation user–target entity pair, we used the top embedding model—identified in the previous stage of the pipeline—to retrieve the user's top five posts supporting the target and the top five opposing it. The user's top stance-relevant posts were concatenated and used as prompt context for LLMs, which were then tasked with labeling the user's stance toward the target entity, following the annotation guidelines outlined in Table 5. In addition to assigning a stance label, the LLMs were instructed to explain their reasoning and identify three specific posts, highlighting the relevant text spans that informed their decision. These explainable outputs facilitate the error analysis presented in section 5.1. The LLM achieving the highest classification performance was ultimately selected to label the users stances toward both target entities on the full dataset. The prompt template used for this task is provided in Appendix C.

4.2.4 Stance Detection on Full Dataset

We employed the pipeline, configured using the validation set, to label the stances of dataset users—excluding the 5% held out for validation—toward both target entities: Trump and Harris. For each user–target entity pair, we first retrieved the user's top stance-relevant posts related

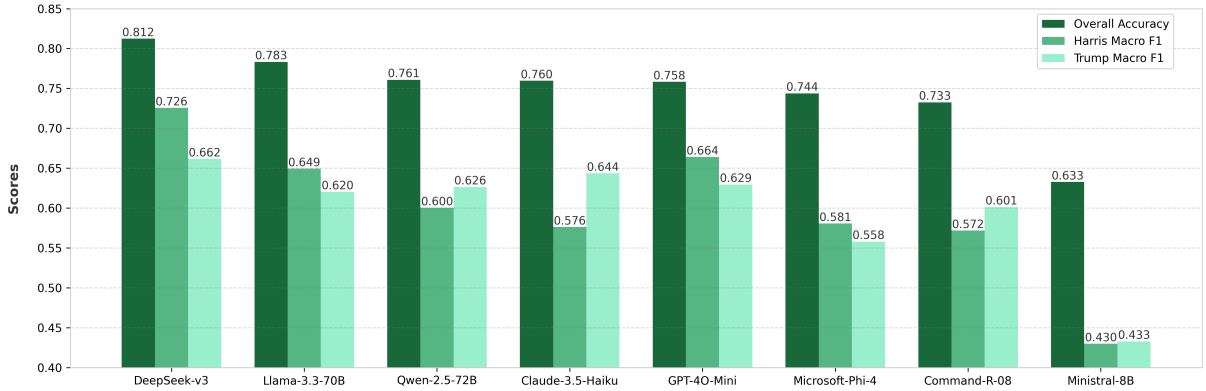

Figure 4: Performance comparison of different language models on stance detection task. The plot shows overall accuracy and entity-specific macro F1 scores for Trump and Harris. Models are ordered by overall accuracy from highest to lowest, demonstrating the relative effectiveness of each model in stance classification.

to the target entity using the best-performing embedding model identified in Section 4.2.2. These posts, along with the corresponding target entity, were then provided as input to the best-performing LLM identified in Section 4.2.3. The LLM generated the user's stance label toward the target entity, along with supporting reasoning and relevant text spans extracted from the user's posts, following the prompt presented in Appendix C. Further details about the structure of the labeled dataset are provided in Appendix D.2.3.

5 Analysis

This section presents a performance analysis of our stance labeling pipeline's components (Figure 3(b)) and error patterns. We conclude by examining the stance distribution characteristics on the final dataset.

5.1 Pipeline Analysis

Embedding Model Performance. In this section, we compare the performance of several state-of-the-art text embedding models in conjunction with *BM25* on our curated validation stance relevancy dataset (See Figure 3(b)). As shown in Figure 5, the *KaLM-mini-v1.5* model (Hu et al., 2025) achieves the highest performance among all the models compared. This result is expected, as the study by Hu et al. (2025) demonstrates the superior performance of this new model relative to others. Specifically, the *KaLM-mini-v1.5* model outperforms the other models in terms of Precision@10 and Recall@10, while its performance is comparable to that of several other models regarding Precision@25 and Recall@25. Given its high performance in retrieving the top 10 documents, which aligns with the configuration used in our stance detection with LLM setup, we have selected this model for stance-based document retrieval.

LLM Performance. We present a systematic evaluation of state-of-the-art large language models (LLMs), with emphasis on computationally efficient options suitable for responsible, large-scale deployment. Our assessment reveals a clear performance hierarchy among the tested models, as illustrated in Figure 4.

DeepSeek-Chat-v3 (Liu et al., 2024) demonstrates exceptional capabilities, achieving the highest overall accuracy of 81.2% and leading macro F1 scores for entity-specific stance detection (66.2% for Trump and 72.6% for Harris). This recent model delivers performance comparable to premium proprietary alternatives while maintaining computational efficiency. *LLAMA-3.3-70B-Instruct* (Dubey et al., 2024) ranks second with 78.3% overall accuracy, followed by *Microsoft's PHI-4* (Abdin et al., 2024) at 74.4%.

Based on these evaluation results, we select *DeepSeek-Chat-v3* as our primary model for the comprehensive stance labeling task due to its superior performance across all metrics.

Stance Confusion Analysis. Examining model performance through the confusion matrix presented in Figure 8, we identify specific classification challenges. The model demonstrates robust capability in identifying oppositional stances, with 91.4% accuracy for the "Against" class. Performance for the "Favor" class remains adequate at 75.7% accuracy. However, substantial difficulties emerge with the "Neither" class, where accuracy drops to 66.8%.

The predominant error pattern involves misclas-

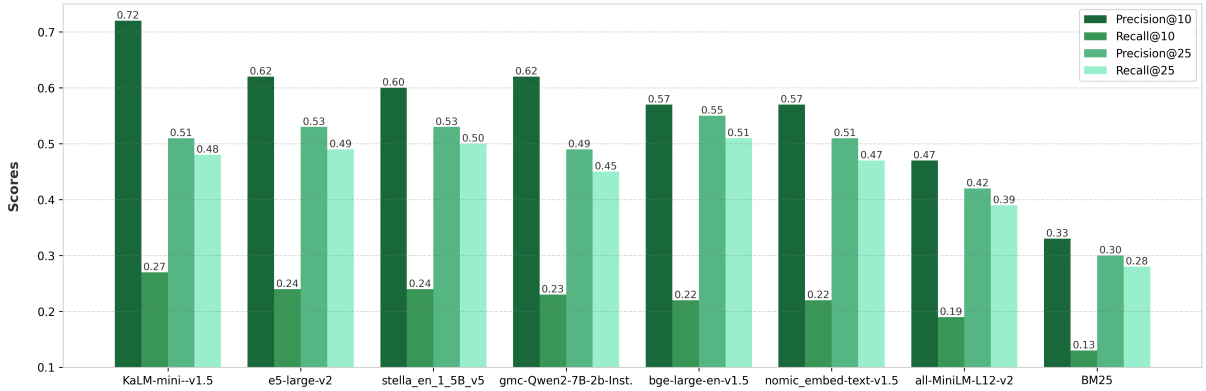

Figure 5: Performance comparison of various text embedding models for stance-based document retrieval.

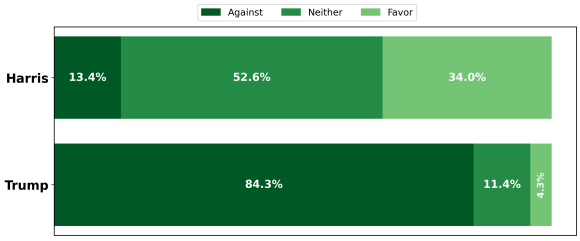

Figure 6: Distribution of final users' stances, labeled by LLM, toward Trump and Harris.

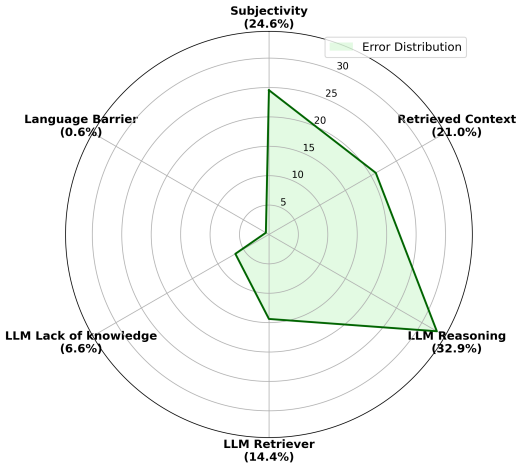

Figure 7: Distribution of reasons for discrepancies between the LLM's predictions and the validation set labels, as detailed in the text.

sification of "Neither" instances as "Favor," occurring in 19.2% of cases. This error typically manifests when users reference target entities within contexts containing positive language or report favorable news without personally expressing support. For example, when users neutrally quote a candidate's statements or report on campaign activities without evaluative commentary, the model sometimes misinterprets these contextual positive signals as stance indicators. This systematic error highlights the nuanced challenge of differentiating between genuine stance expression and neutral reporting that contains positive linguistic elements—a distinction that requires sophisticated contextual understanding beyond surface sentiment analysis.

Error Categorization. To understand discrepancies between LLM predictions and our validation labels, we extracted and analyzed misclassified instances, identifying six distinct error categories as illustrated in Figure 7:

Context: Insufficient quality of retrieved posts, preventing accurate stance determination.

Reasoning: Flawed logical processing by the LLM, particularly when confronted with sarcasm or irony.

Subjectivity: Posts allowing multiple valid interpretations. For example, "I will vote for Harris. I hate Harris, but we can deal with her government a lot easier than Trump's government" was classified by the LLM as favorable to Harris (based on voting intention), while our validation labeled it "neither."

LLM retriever: Failure to ground reasoning in relevant evidence despite its presence in retrieved content. For instance, overlooking explicit declarations like "I voted for her #Harris2024" when making stance determinations.

LLM's lack of knowledge: Inability to recognize platform-specific political signifiers, such as emoji combinations (blue heart + wave emoji signifying "blue wave" or blue heart + ballot box representing #VoteBlue) that indicate Democratic Party support.

Language barrier: Content in non-English languages leading to misinterpretation (observed in only one case).

Our findings reveals important insights about stance detection challenges. Nearly 80% of er-

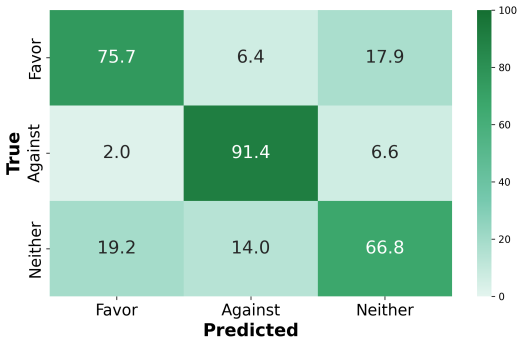

Figure 8: Confusion matrix showing the performance of stance detection across three classes: Favor, Against, and Neither. Values are shown as percentages of true labels against our validation dataset.

rors stem from three categories: reasoning limitations (32.9%), inherent subjectivity (24.6%), and context retrieval issues (21.0%). Approximately 35% of errors relate to potentially addressable technical implementation issues (Context + LLM retriever), while 65% involve fundamental challenges in language understanding and interpretation. This suggests that improvements to stance detection systems should focus on both enhancing retrieval mechanisms and developing more sophisticated approaches to reasoning about political discourse, particularly for handling ambiguity and platform-specific communication conventions.

5.2 Stance Distribution Characteristics

Our analysis of stance distributions within the labeled dataset reveals distinct patterns regarding user attitudes toward the two candidates. As shown in Figure 6, a majority of users express opposition toward Trump, aligning with previous research on Bluesky's left-leaning demographic characteristics (Failla and Rossetti, 2024). Conversely, most users do not express opposition toward Harris—though this absence of negative sentiment should not be interpreted as explicit support, but rather as neutral or non-oppositional positioning.

6 Future Outlook and Broader Impact

In an era where declining data accessibility from major social platforms has created significant obstacles for researchers, our work offers both a timely resource and a replicable methodological blueprint. The public release of POLITISKY24 provides a foundational, open-data cornerstone intended to catalyze a new wave of transparent and verifiable research on emerging platforms like Bluesky. Beyond the dataset itself, the robust pipeline developed for its creation serves as an adaptable framework for the community, demonstrating a clear path forward for creating high-quality, labeled datasets from other open platforms.

The dataset's primary value lies in the new interdisciplinary research it enables. While annotated for stance, POLITISKY24 is designed as a foundational layer that can be extended to investigate a wide range of complex social and political phenomena. Researchers can introduce new annotations to study the intricate interplay between a user's political stance and their online behaviors, such as the dissemination of misinformation, the use of hyper-partisan language, or engagement with hate speech. Furthermore, the dataset invites a powerful integration of its stance labels with its rich network and temporal data. By analyzing the network topology, researchers can map the structure of political echo chambers, identify influential actors who drive conversations, and model the diffusion of specific narratives through the platform. The longitudinal nature of the data also allows for dynamic, event-driven analysis, enabling studies that track how collective opinion shifts in response to major real-world events, such as candidate debates, policy announcements, or breaking news.

7 Conclusion

In this study, we constructed a labeled dataset capturing user stances on the Bluesky social network toward the main 2024 U.S. presidential candidates: Trump and Harris. Labels were assigned using a pipeline that combined text embedding models for context retrieval with a large language model (LLM) for stance detection, achieving an overall accuracy of 81%. Among the 19% of misclassifications, roughly one-third stemmed from context + LLM retriever errors, while the remaining two-thirds were attributed to challenges in language understanding and interpretation. Among the models tested, the *KaLM-mini-v1.5* embedding model and *DeepSeek-v3* LLM outperformed other state-of-the-art alternatives in this pipeline.

For future work, we aim to leverage users' graph networks to assist with data labeling. In addition, as a separate direction, we plan to apply our pipeline across different time windows to examine how users' stances evolve over time.

Limitations

It is important to acknowledge certain considerations regarding our approach and dataset:

First, our methodology for selecting users for the validation set, which drew upon established practices of hashtag-based sampling found in existing literature, while practical, may introduce sampling biases. Users who explicitly utilize hashtags could represent a particular subset of the Bluesky user population, potentially exhibiting more extreme or clearly articulated stances.

Second, beyond the limitations of the hashtag-based sampling method, another concern is the left-leaning bias among Bluesky users, which leads to an imbalanced validation set. As shown in Figure 13, approximately 87.6% of users in the validation set oppose Trump, while only 3.6% support him. Regarding Harris's target group, about 11.2% are against her, whereas 49.2% are in favor.

Third, while we employed detailed annotation guidelines and measured inter-rater agreement among expert annotators to minimize subjectivity, the inherent nature of human judgment in interpreting nuanced stance expressions remains a consideration. Stance often contains implicit signals that can be interpreted differently, even by trained annotators.

Fourth, resource constraints necessitated a selection of LLMs for evaluation in our experimentation pipeline. While we aimed to include models representing diverse architectures and parameter sizes, our findings concerning model performance for labeling may not comprehensively generalize across the entire spectrum of available language models, especially given the rapid emergence of new models.

Finally, while the pipeline is designed to be robust and adaptable, applying it to other platforms would require a few adjustments. Features such as hashtags and user post histories are generally transferable, but platform-specific components—particularly the seed user selection step—would need to be redefined (e.g., using trending hashtags instead of Bluesky's custom feeds).

Ethical Considerations

This research was conducted in strict adherence to established ethical guidelines and Bluesky's terms of service and data use policies[2], emphasizing user privacy within the platform's open data context. Bluesky's commitment to transparency facilitates research, and our practices align accordingly. Data was collected exclusively through official APIs and libraries, ensuring compliance and implementing data minimization techniques to gather only information essential for this study. While user identifiers (such as DIDs or handles) and post content are inherently public on Bluesky, we have processed and will present the data responsibly. The released dataset—comprising user-target stance pairs, supporting rationales, text spans, and accompanying networks—utilizes this publicly accessible user-generated content. It is intended strictly for research purposes, aiming to facilitate advancements in computational social science. All subsequent uses of this dataset must also adhere to Bluesky's policies. Despite Bluesky's open data model and our commitment to ethical handling, the sensitive nature of political stance information warrants careful consideration. Although sourced from public posts, there is a potential risk that the curated stance labels, if de-contextualized or combined with other information, could be misused for unintended user profiling or contribute to targeted harassment. We urge future users to handle this data with a strong sense of responsibility and full awareness of these potential implications.

[2]https://bsky.social/about/support/privacy-policy

Acknowledgments

This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences, IPM, Iran (No. CS1403-4-05).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Leonhard Balduf, Saidu Sokoto, Onur Ascigil, Gareth Tyson, Björn Scheuermann, Maciej Korczyński, Ignacio Castro, and Michał Król. 2024. Looking AT the blue skies of bluesky. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pages 76–91, New York, NY, USA. ACM.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 141–152.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Subhabrata Dutta, Samiya Caur, Soumen Chakrabarti, and Tanmoy Chakraborty. 2022. Semi-supervised stance detection of tweets via distant network supervision. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 241–251.

Andrea Failla and Giulio Rossetti. 2024. "i'm in the bluesky tonight": Insights from a year worth of social data. *PLoS One*, 19(11):e0310330.

Marilyn Fernandez and Laura Nichols. 2002. Bridging and bonding capital: Pluralist ethnic relations in silicon valley. *International Journal of Sociology and Social Policy*, 22:104–122.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, and 1 others. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Ujun Jeong, Bohan Jiang, Zhen Tan, H Russell Bernard, and Huan Liu. 2024. Descriptor: A temporal multi-network dataset of social interactions in bluesky social (BlueTempNet). *IEEE Data Descr.*, 1:71–79.

Kornraphop Kawintiranon and Lisa Singh. 2021a. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021b. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Martin Kleppmann, Paul Frazee, Jake Gold, Jay Graber, Daniel Holmgren, Devin Ivy, Jeromy Johnson, Bryan Newbold, and Jaz Volpert. 2024. Bluesky and the AT protocol: Usable decentralized social media. In *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet*, pages 1–7, New York, NY, USA. ACM.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021b. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. A dataset for detecting stance in tweets. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 3945–3952.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Dorian Quelle and Alexandre Bovet. 2024. Bluesky: Network topology, polarisation, and algorithmic curation. *arXiv preprint arXiv:2405.17571*.

Dorian Quelle and Alexandre Bovet. 2025. Bluesky: Network topology, polarization, and algorithmic curation. *PloS one*, 20(2):e0318034.

Vahid Rahimzadeh, Ali Hamzehpour, Azadeh Shakery, and Masoud Asadpour. 2025. From millions of tweets to actionable insights: Leveraging llms for user profiling. *Preprint*, arXiv:2505.06184.

Majid Zarharan, Maryam Hashemi, Malika Behroozrazegh, Sauleh Eetemadi, Mohammad Taher Pilehvar, and Jennifer Foster. 2024. Farexstance: Explainable stance detection for farsi. *arXiv preprint arXiv:2412.14008*.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, and Hu Huang. 2024a. A survey of stance detection on social media: New directions and perspectives. *arXiv preprint arXiv:2409.15690*.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2024b. Doubleh: Twitter user stance detection via bipartite graph neural networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1766–1778.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics.

A Experimental Setups

All LLM inference for stance detection was performed using OpenRouter (https://openrouter.ai/) with temperature set to 1e-10 to ensure deterministic outputs. Embedding model evaluation experiments were conducted using Google Colab environments equipped with NVIDIA T4 GPUs. For the data embedding process, we utilized NVIDIA RTX 6000 GPUs to efficiently process the large volume of user content in our dataset.

B Annotation Guidelines

The annotation guidelines used to decide on stance of the users toward targets are given in Table 5. Our annotation process was supported by three male experts in U.S. Politics, aged 26, 27, and 32, who provided specialized knowledge essential for accurate stance labeling.

C Prompts

The prompt used for stance labeling using LLMs is shown in Figure 9.

Stance label	Description
Favor	• Directly expressing support for the target entity. • Expressing support for the target's political party in the 2024 elections.
Against	• Directly expressing opposition to the target entity. • Expressing opposition to the target's political party in the 2024 elections.
Neither	• Expressing no clear stance or a mixture of support and opposition. • Sharing news without expressing a personal opinion.

Table 5: Annotation guidelines for stance labeling.

D Dataset Details

In this section, we provide additional details about the U.S. political Bluesky dataset we collected, as well as the stance dataset we built upon.

D.1 Bluesky dataset

The collected Bluesky dataset consists of three main components: feeds, user post histories, and the user network. In the following subsections, we examine each of these components in detail.

D.1.1 Feeds

Table 6 presents a summary of the statistics for the feeds used to collect data during the period from November 12 to 27, 2024. In total, 8,561 unique users—each with at least 10 posts across these three feeds—were included in our dataset.

D.1.2 User Post History

Users' post histories are stored using the fields specified in Table 7, which lists each field along with its valid values and detailed descriptions.

In the following, we examine the distribution of posts and active users over time, as well as the distribution of posts per user and the lengths of those posts. Figure 10 illustrates the temporal distribution of posts (including both original posts and reposts) published by the 8,561 selected unique users. As shown, the number of posts increased over time, with a significant surge in November 2024, the final month of data collection. This surge coincides with the U.S. election month, making these recent posts particularly valuable for analyz-

```
SYSTEM_PROMPT = """
    You are a researcher for social network analysis in computer science and this task is meant for a paper to be submitted to ICWSM 2025.
    You are given a target entity alongside a set of relevant user tweets for the task. This is about US Presidential Election 2024 between Donald Trump and
    Kamala Harris.
    The task is to find the stance of user regarding the target entity.
    Keep in mind that trump has won the election and tweets are gatehered in both before and after the election.

    The valid stances are:
    - Favor
        When user is in favor of the target entity directly. Another scenario is where the user is in favor of the target's party or policies with election in mind.
    - Against
        When user is against the target entity directly. Another scenario is where the user is against the target's party or policies with election in mind.
    - Neither
        If the user is not taking any stance or has both tweets in favor of and against the target entity in a balance way. Keep in mind that we want the overall
 stance and users also may be sarcastic, so if the user have some critical opinions containing stance and in the meantime have some mild opinions,
 you should not label it as Neither

    IMPORTANT: You must strictly follow these output format rules:
    1. Use double quotes for all JSON keys and string values
    2. Include all required fields
    3. Use exactly these stance values: "Favor", "Against", "Neither"
    4. Ensure the confidence value is a number between 0 and 1 (not a string)
    5. Do not add any explanatory text before or after the JSON
    6. Ensure arrays are properly formatted with square brackets and commas

    Take these steps:
    1. Find the most relevant tweets and spans for determining the stance
    2. If you can't find any relevant tweet, the stance would be Neither
    3. If you find any relevant tweet, try to determine whether the user stance is against or in favor of the target entity

    For unrelated tweets, output exactly this format:
    {{
        "stance": "Neither",
        "reason": "your reasoning here"
    }}

    For related tweets, your output should consist of the source tweets and spans used for reasoning, and the reasoning itself for the stance detection, alongside
     with the stance itself.
    Output format MUST be exactly like below:
    {{
        "source_tweets": ["tweet1 used for reasoning", "tweet2 used for reasoning", "tweet3 used for reasoning"],
        "spans": ["span1 used for reasoning", "span2 used for reasoning", "span3 used for reasoning"],
        "reason": "your reasoning here",
        "stance": "Against/Favor",
        "confidence": "your confidence between 0 and 1"
    }}

    Remember: Your response must contain ONLY the JSON object, nothing else before or after it. You MUST take the given steps to determine the stance.
    """
USER_PROMPT = "Target Entity: {target_entity} \nRelevant Tweets: {relevant_tweets}"
```

Figure 9: The prompt used for stance labeling using LLMs in the last stage of our stance labeling pipeline.

Feed Name	Feed URL	# Feed Posts	# Feed Users	# Selected Users
US Politics	https://bsky.app/profile/did:plc:7mtqkeetxgxqfyhyi2dnyga2/feed/aaadhh6hwvaca	325,065	137,137	8,139
PolSky	https://bsky.app/profile/did:plc:cmqylb7cttgdivvolbnyoxui/feed/aaabr7dsksuvw	251,112	113,794	8,311
Unknown*	https://bsky.app/profile/did:plc:jw2yabtwnjmyi3q7vqrhxn7a/feed/aaamfsfbh26y6	20,593	13,974	2,800

* This feed is currently unavailable from its creator.

Table 6: Overview of feed statistics.

ing public stances toward the election candidates. A similar pattern is observed in Figure 11, which presents the temporal distribution of active users (defined as those with at least 10 posts in each time period).

Figure 12 illustrates the distribution of Bluesky post lengths, measured by word count. It shows an inverse relationship between post length and frequency, with longer posts occurring less frequently. Notably, the majority of posts contain fewer than 60 words.

D.1.3 User Network

Table 8 presents the schema of the user network dataset, which captures interactions between users

Field	Value	Description
Post Id	Post identifier	A unique identifier assigned to the post.
User Id	User identifier	A unique identifier of the user who published the post.
Post Time	Post timestamp	The timestamp when the post was published.
Content	Post content	The text content of the post.
Hashtags	List of hashtags	Hashtags included in the post content.
Languages	List of languages	Languages used in the post content.
Mentions	List of mentions	Users mentioned in the post.
Links	List of URLs	Hyperlinks included in the post.
Is Repost	True / False	Indicates whether the post is a repost of another post.
Source User Id	Source user identifier	The original user identifier if the post is a repost.
Is Quote	True / False	Indicates whether the post contains a quote of another post.
Quote Id	Quoted post identifier	The identifier of the quoted post, if any.
Quote User Id	Quoted user identifier	The identifier of the author of the quoted post.
Quote Content	Quote content	The text content of the quoted post.
Is Reply	True / False	Indicates whether the post is a reply to another post.
Parent Id	Parent post identifier	The identifier of the immediate parent post (if replying).
Parent User Id	Parent user identifier	The identifier of the user who authored the parent post.
Root Id	Root post identifier	The identifier of the root post in a reply chain (if replying).
Root User Id	Root user identifier	The identifier of the user who authored the root post.

Table 7: Schema of the user post history dataset, detailing the fields, possible values, and their descriptions.

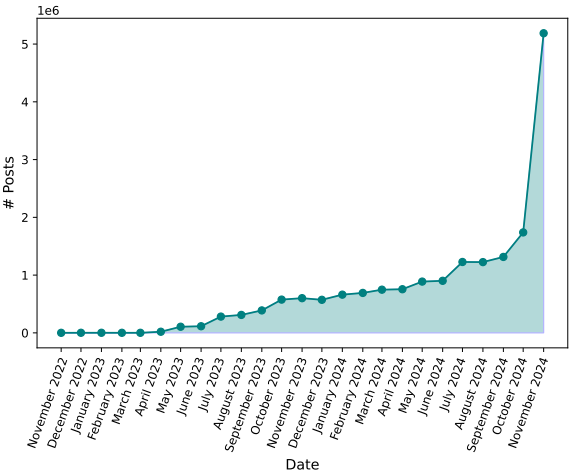

Figure 10: Post distribution over time.

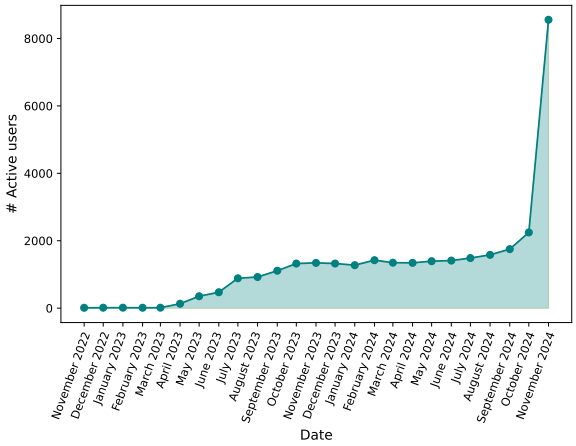

Figure 11: Active user distribution over time. 'Active' refers to users who published or reposted at least 10 posts during each respective time period.

through likes and reposts.

D.2 Stance dataset

The annotated stance dataset comprises three key components: 1. stance relevancy (validation dataset), 2. user stance (validation dataset), and 3. user stance (full dataset). In the following subsections, each of these components will be analyzed in detail.

Field	Value	Description
Source User Id	Source user identifier	The identifier of the source user.
Target User Id	Target user identifier	The identifier of the target user.
Interaction type	Repost / Like	The type of interaction from the source user toward the target user.
Count	Interaction count	The number of times the source user interacted with the target user.

Table 8: User network dataset schema.

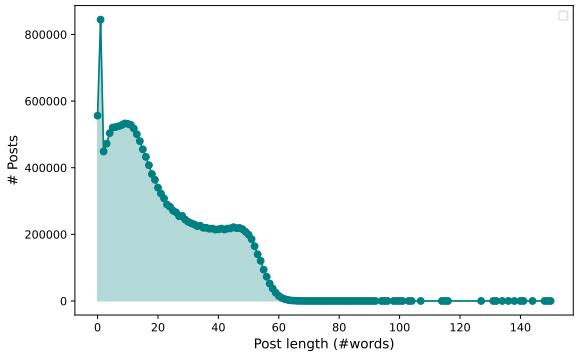

Figure 12: Word count distribution of the Bluesky posts.

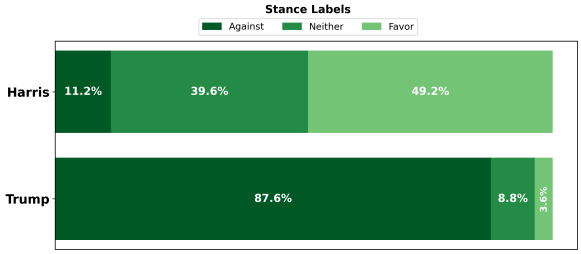

Figure 13: Distribution of validation users' stances, labeled by human experts, toward Trump and Harris.

D.2.1 Stance Relevancy (Validation dataset)

As described in Section 4.2.2, the construction of the validation stance relevancy dataset involved two main steps. First, a subset of posts was manually annotated with stance labels toward Trump and Harris. The structure of this human-annotated dataset is presented in Table 9, and summary statistics of the labeled posts are provided in Table 10. As shown, at least 25 posts were annotated for each combination of target and stance category. In the second step, a query–post stance relevancy dataset was derived from the manually annotated data. The structure of this derived dataset is presented in Table 11.

D.2.2 User Stance (Validation dataset)

Table 12 presents the schema of the validation user stance dataset. Table 13 reports statistics on the number of users labeled by humans, the LLM, and both, categorized by stance and target.

Figure 13 shows the distribution of labels toward our target entities in the human-annotated user stance validation set. As illustrated, the stance labels indicate strong opposition to Trump and mild opposition to Harris. This pattern aligns with the left-leaning political orientation of the Bluesky community (Quelle and Bovet, 2024).

D.2.3 User Stance (Full dataset)

The schema of the user stance dataset, annotated by the LLM, is presented in Table 14. This table details all dataset fields, including their possible values and descriptions. Additionally, Table 15 provides statistics on the context posts and LLM-generated fields across the entire dataset.

D.2.4 User Stance Examples

In Table 16, we present examples of dataset users' stances toward Trump and Harris.

E LLM Usage Statement

We acknowledge the use of Large Language Models solely for grammar checking and paraphrasing. All ideas, analyses, and substantive content presented in this paper are the original work of the authors. LLMs were employed only to enhance readability and conciseness of our contributions.

Field	Value	Description
Post Id	Post identifier	A unique identifier for the post
Trump	Favor / Against / Neither	Stance label expressed in the post toward Trump
Harris	Favor / Against / Neither	Stance label expressed in the post toward Harris

Table 9: Schema of the human-annotated validation stance relevancy dataset (post–target entity pairs).

Target Entity	Stance Label	# Human-Annotated Posts
Trump	Favor	25
	Against	36
	Neither	114
Harris	Favor	25
	Against	26
	Neither	124

Table 10: Number of human-annotated posts for each stance toward Trump and Harris.

Field	Value	Description
Query	Query content	A query expressing a supportive or opposing stance toward a given target (Trump / Harris).
Post Id	Post identifier	The unique identifier of the post.
Stance Relevance Label	Relevant / Non-relevant	Indicates whether the post is relevant to the query in terms of both target and stance.

Table 11: Schema of the human-annotated validation stance relevancy (query–post stance relevancy pairs).

Field	Value	Description
User Id	User identifier	A unique identifier for the user
Trump	Favor / Against / Neither	Overall stance label expressed in the user's post history toward Trump.
Harris	Favor / Against / Neither	Overall stance label expressed in the user's post history toward Harris.

Table 12: Schema of the human-annotated validation user stance dataset.

Target Entity	Stance Label	# Human-Annotated Users	# LLM-Annotated Users	# Common Users
Trump	Favor	16	29	14
	Against	390	379	362
	Neither	39	37	16
Harris	Favor	219	199	164
	Against	50	69	40
	Neither	176	177	127

Table 13: Number of human and LLM-annotated users for each stance toward Trump and Harris.

Field	Value	Description
User Id	User identifier	A unique identifier for the user.
Target Entity	Trump / Harris	The entity toward which the user's stance is labeled.
Context	List of retrieved posts	A collection of posts retrieved by the embedding model.
Source Posts	List of source posts	A subset of posts selected by the LLM from the context.
Spans	List of text spans	Specific portions of the source posts used by the LLM for labeling the stance.
Reason	Reason content	The rationale provided for choosing the stance label.
Stance Label	Favor / Against / Neither	The stance classification toward the target entity.
Confidence Level	Real number between 0 and 1	A numerical value representing the confidence in the stance classification.

Table 14: Schema of the LLM-annotated user stance dataset.

Target Entity	Stance Label	# LLM-Annotated Users	Avg #Context Posts	Avg #Source Posts	Avg Confidence Level
Trump	Favor	345	7.01	3.71	0.89
	Against	6,759	7.39	4.13	0.94
	Neither	918	6.74	2.29	0.07
Harris	Favor	2,726	7.62	3.26	0.91
	Against	1,073	7.71	2.85	0.88
	Neither	4,223	7.67	2.05	0.02

Table 15: Summary of statistics generated by the LLM for labeling user stances in the full dataset.

User stance labels	Relevant posts	
	Trump target	Harris target
Against Trump, Against Harris	"...Welcome to the unproductive chaos of Trump..."	"...I also think that when Oprah was campaigning for Harris the quasi singing of Kamala was - was – like fingernails on a chalkboard. . . "
Against Trump, Favor Harris	"...Making America healthy ? Really ?? I think not !!!..."	"...May have been the losing side. Still not convinced it was the wrong one."\n\n#firefly #serenity #election2024 #HarrisWalz #democracy #politics #antiMAGA #justice #resist..."
Favor Trump, Against Harris	"...God bless president Trump and the United States of america!..."	"...$1 billion disaster': Here's what FEC filings show about Harris campaign's 3 month spending spree. . . "

Table 16: Examples of dataset users' stances toward Trump and Harris.