Bias Detection in Political Social Media

Xingrui Huang
Northeastern University
huang.xingr@northeastern.edu

Liangyou Xu
Northeastern University
xu.lian@northeastern.edu

Abstract

Detecting political bias in social media remains challenging due to informal language and platform-specific conventions. We address this through a cross-platform study of Truth Social and Bluesky, platforms with contrasting ideological compositions. We develop an LLM-assisted labeling pipeline that directly assigns Left/Neutral/Right labels to posts and construct a topic-annotated corpus of approximately 130,000 annotated posts.

Benchmarking shows DistilBERT achieves 75% accuracy with balanced per-class performance, substantially outperforming Naive Bayes, particularly on Neutral content (F1 from 0.643 to 0.756). Platform analysis reveals Truth Social's ideological homogeneity and bimodal sentiment versus Bluesky's diversity and neutral tone. Semantic clustering identifies distinct specialization: Truth Social centers on political identity and institutional skepticism, while Bluesky emphasizes creative expression and multilingual interaction.

Temporal validation exposes model limitations. Applied to Congressional tweets (2008–2017), the classifier shows apparent polarization trends as perceived by the model but also exhibits systematic biases. On the 2005 Politics.com forum, accuracy drops to 17.4%, revealing poor calibration on self-identified affiliations. Our work contributes a reusable cross-platform corpus, strong baseline models, and diagnostic tools for studying both online polarization and the limitations of bias classifiers under domain shift.

Keywords

political bias detection; stance; framing; polarization; social media; Twitter/X; Truth Social; Bluesky

1 Introduction

The increasing ideological polarization of online discourse has transformed social media into both a mirror and a catalyst of political division. While traditional journalism has long been studied for partisan bias, social platforms now play a more dynamic role in shaping political narratives. Users not only consume but also produce politically charged content, often within ideologically homogeneous environments that reinforce confirmation bias. Platforms such as Truth Social and Bluesky exemplify this polarization, each fostering distinct linguistic norms, framing strategies, and emotional tones that reflect and amplify their users' ideological orientations. Detecting and quantifying political bias in such user-generated text is therefore essential for understanding online polarization and developing computational tools that can characterize the language of ideology at scale.

However, identifying bias in short, informal, and context-dependent social-media posts remains a formidable challenge. Compared with news articles, social-media text lacks consistent structure, contains sarcasm and slang, and often embeds ideological stance through framing rather than explicit declaration. Moreover, reliable large-scale datasets covering multiple platforms are scarce; most prior corpora are derived from Twitter or news media, limiting the generality of bias-detection models. Addressing these challenges requires both better data collection methods and more robust modeling approaches that can handle the linguistic diversity of emerging social networks. To bridge this gap, we propose a cross-platform study that develops a dataset of approximately 130,000 posts collected from Truth Social and Bluesky, annotated along a Left/Neutral/Right bias spectrum. These platforms were selected due to their contrasting ideological compositions, with Truth Social demonstrating high ideological homogeneity and Bluesky showing greater ideological diversity. Our dataset uses an LLM-assisted annotation pipeline rather than hand-engineered weak supervision for labeling. Posts are collected via keyword search across twenty political topics (e.g., abortion, immigration, climate), then each post is independently labeled by prompting a large language model to assign Left/Neutral/Right based solely on text content, without multi-stage refinement or hashtag heuristics. We use keyword framing only to validate label quality post-hoc, not as supervision signal. In addition, we use a simple platform-level noisy prior (labeling Bluesky as left-leaning and Truth Social as right-leaning) as a comparison baseline and as an intentionally noisy supervision regime, rather than as a source of refined supervision for the LLM labels themselves.

Building on this dataset, we benchmark a range of models from interpretable linear baselines to transformer architectures. Our results demonstrate that contextualized models, such as DistilBERT, substantially outperform classical baselines (e.g., Naive Bayes), achieving an approximate 10 percentage point gain in Macro F1 and significantly improving the classification of Neutral content. Furthermore, our evaluation reveals pronounced platform-specific linguistic artifacts, such as Truth Social's highly bimodal, U-shaped sentiment distribution contrasted with Bluesky's dominant neutral sentiment spike. We apply cross-platform semantic clustering to map the joint discourse space, which reveals key thematic specialization: Truth Social clusters coalesce around political identity and institutional skepticism, while Bluesky centers on creative expression and multilingual interaction. Finally, we validate our bias models by applying them to the US Congressional Tweets Dataset (2008–2017), which suggests that the models capture meaningful ideological signals by tracking an apparent acceleration of polarization and a decline of neutral content in elite political discourse, while also reflecting their own calibration biases.

We organize our investigation around four research questions that span data quality, model architecture, linguistic feature analysis, and domain generalization. We then benchmark classical and transformer models and apply the resulting classifier to temporally and stylistically out-of-domain corpora to evaluate whether linguistic signals of polarization remain consistent across platforms and over time.

Research Questions. We investigate four questions in our study:

1. How can we effectively model and quantify political bias in short, user-generated text from ideologically distinct platforms?
2. Which linguistic features such as stance, framing, and sentiment are most indicative of partisan bias, and how do these features vary across platforms and topics?
3. How robust are bias classifiers trained on contemporary partisan platforms when applied out-of-domain across platforms and over time (e.g., to US Congressional tweets and legacy forums)?
4. What semantic and topical structure emerges when we map cross-platform discourse in a shared embedding space?

Contributions.

1. **Dataset.** Novel Cross-Platform Corpus and Characterization: Release of a topic-platform-annotated corpus of 131,808 posts from Truth Social and Bluesky (93,614 from Bluesky and 38,194 from Truth Social) with Left/Neutral/Right bias labels, systematically characterizing the ideological homogeneity of Truth Social and the diversity of Bluesky.
2. **Dual-label experimental design.** LLM-generated Left/Neutral/Right annotations alongside platform-as-label baselines (all Bluesky as Left, all Truth Social as Right). Comparing models trained on these two label regimes reveals that LLM labels yield stronger text-only classifiers (85.8% vs 81.2% accuracy), but platform identity alone achieves 83.8% accuracy on the L/R subset, highlighting a major confound in cross-platform bias detection.
3. **Cross-Platform Benchmarking.** Establishment of reproducible baselines showing that contextualized transformers (DistilBERT) substantially outperform classical models, especially in distinguishing neutral content from partisan framing (improving Neutral F1 from 0.643 to 0.756).
4. **Sociolinguistic Insights via Semantic Mapping.** Comprehensive semantic clustering analysis that maps the platforms' specialized content domains, while identifying shared discourse themes that transcend ideological boundaries.
5. **Temporal Analysis of Polarization.** Illustration of how the trained bias classifier perceives partisan polarization in longitudinal data (US Congressional Tweets, 2008–2017), coupled with analysis of where these trends reflect genuine temporal change versus model-induced bias.

2 Related Work

Research on political bias detection has evolved along two parallel tracks: studies focused on traditional news media where bias manifests through editorial choices and framing, and more recent work on social media where informal language and user context dominate. We review both lines of inquiry and highlight how methodological insights from each inform our cross-platform approach.

2.1 Bias Detection in News Media

Political bias detection has been extensively studied in the context of traditional news, where language follows formal and editorial conventions. Early research framed bias detection as a text classification problem, using supervised learning with handcrafted features. Classical models such as logistic regression and support vector machines, trained on term frequency representations, demonstrated that lexical and syntactic patterns can reveal ideological leanings [19]. These methods exploited sentiment polarity, framing verbs, and the frequency of partisan cues to infer the political orientation of news outlets or articles.

The field evolved rapidly with the rise of deep contextualized representations. The introduction of large pretrained transformer models fundamentally improved the modeling of subtle ideological cues by capturing long-range dependencies and contextual semantics. BERT and its variants enabled systems to distinguish between statements that differ only in framing or word choice [5]. Subsequent studies extended these architectures to explicitly predict political ideology or hyperpartisanship. For instance, Baly and colleagues leveraged transformer-based models to infer the ideological slant of news articles and outlets, while the SemEval-2019 Hyperpartisan News Detection task provided a benchmark for distinguishing between partisan and neutral writing styles [1, 11]. These advances established transformer-based methods as the standard for bias detection in structured, long-form text.

However, despite their success, these models are trained and evaluated on well-edited corpora where bias is expressed through framing and selection rather than slang or sarcasm. As a result, they struggle to generalize to the noisy, unstructured, and often emotionally charged text found on social media. The methodological lessons from news bias detection, especially the importance of feature interpretability and contextual modeling, serve as a foundation but also highlight the domain gap that motivates our work.

2.2 Bias Detection in Social Media

Detecting political bias in social media introduces new challenges due to brevity, informality, and the strong role of user context. Posts are often emotionally charged, rich in irony, and embedded in platform-specific conventions. Much of the literature has therefore addressed adjacent problems such as stance detection, misinformation classification, and topic-dependent framing rather than generalized ideological classification. The SemEval-2016 stance detection shared task remains a cornerstone dataset and has frequently been used as a transfer source for political natural language processing beyond sentiment analysis [14].

Large-scale pretrained encoders have substantially improved modeling of informal text. The TweetEval benchmark established that transformers can adapt effectively to short, noisy inputs across sentiment, stance, and toxicity tasks [2]. Domain-specific pretraining extends this capability further. PoliBERTweet, trained on tens of millions of U.S. election tweets, consistently outperforms generic encoders on political Twitter tasks, indicating that alignment between pretraining data and target domain matters as much as raw model size [10].

Beyond text-only signals, hybrid approaches that couple language with diffusion structure have shown promise. Retweet-BERT, which fuses transformer representations with retweet topology, achieves strong performance on COVID-19 and 2020 election data and surfaces asymmetric echo-chamber patterns [9]. These results suggest that linguistic cues, social ties, and amplification dynamics each carry partisan signal and can be profitably combined.

The scope of social media bias research has also broadened beyond binary left-right classification. Work on fine-grained ideology prediction shows that moderates and neutrals are identifiable when labels are available at scale, but annotation costs rise sharply [16]. This reality motivates weak and semi-supervised labeling strategies. Stance resources such as P-Stance complement the original SemEval dataset and enable cross-target generalization [13]. Adversarial domain adaptation demonstrates that bias-relevant representations can transfer from news to social media without overfitting to the source domain [1].

Multilingual and cross-lingual research further demonstrates portability. Stylistic and affective features aid transfer across European languages [12]. Large language model-aided zero-shot methods can elicit and retrieve stance knowledge for low-resource targets [22], though task-specific fine-tuned encoders often still outperform large generative models in non-English settings [20]. Unsupervised approaches can also recover meaningful latent ideology directly from social traces when labels are scarce [6].

Empirically, the field now benefits from very large corpora and coverage of new platforms. The Election2020 collection contains over a billion tweets documenting election discourse [3]. Reddit datasets span multi-axis ideology and extreme-bias communities, enabling studies of homophily, heterophily, and radicalization [4, 17]. For emerging ecosystems, PolitiSky24 introduces the first stance dataset on Bluesky with user-level labels and interaction graphs [18]. By contrast, peer-reviewed corpora for Truth Social remain scarce, limiting cross-platform generalization studies and motivating our data collection effort.

These observations shape our research design. We build a topic-annotated corpus spanning two ideologically contrasting platforms and use an LLM-assisted annotation pipeline to obtain Left / Neutral / Right labels (Section 4.1). We then benchmark classical and transformer models and deploy a fine-tuned DistilBERT classifier on multiple corpora to evaluate whether linguistic signals of polarization remain consistent across platforms and over time.

3 Datasets

New Corpus. We construct an English-language corpus of 131,808 posts drawn from two ideologically divergent platforms—Truth Social and Bluesky—covering around twenty politically salient topics such as abortion, climate, immigration, gun control, and healthcare. After filtering out posts shorter than five characters and deduplicating by `text` and `post_id`, the final corpus comprises 93,614 Bluesky posts and 38,194 Truth Social posts. Each post is annotated with platform and topic labels to enable both cross-platform and cross-topic evaluation. Political bias labels (*Left, Neutral, Right*) are obtained through an LLM-assisted annotation pipeline (Section 4.1). For each post we prompt Gemini 2.5 Flash to assign a single Left/Neutral/Right label based on the post text, and we treat this LLM output as our primary supervision signal. In addition, we define a simple platform-prior label that maps all Bluesky posts to *Left* and all Truth Social posts to *Right*. These platform-derived labels are used only to train and evaluate noisy baselines and to quantify how predictive platform identity is; they are not used to construct or refine the LLM labels themselves.

Early Platform Data for Cross-Platform Analysis. To examine whether ideological discourse patterns remain consistent across platforms or evolve with platform-specific norms, we construct a balanced comparative corpus drawing from two existing datasets. We sample 80,000 posts from Truth Social [8] collected between February and October 2022, and 80,000 posts from Bluesky [7] collected between February 17th, 2023, and March 18th, 2024 (inclusive). These time windows capture the first year of each platform, enabling comparison of discourse during analogous developmental phases despite the two-year temporal offset.

We apply consistent preprocessing to ensure comparability: excluding reposts and replies to focus on original content, and filtering to retain only English-language posts. This controlled sampling design isolates platform effects from temporal and linguistic confounds. We use this corpus for embedding-based analysis to test whether posts from ideologically similar users cluster together regardless of platform, or whether platform-specific linguistic conventions fragment the ideological space. By projecting posts into a shared semantic embedding space and measuring cross-platform versus within-platform similarity, we can assess whether Truth Social and Bluesky function as echo chambers with distinct linguistic identities, or whether partisan discourse follows consistent patterns that transcend individual platforms. This analysis complements our primary bias detection task by revealing the underlying structure of political language across social media ecosystems. After additionally discarding posts shorter than ten characters, this sampling yields 51,379 Truth Social posts and 64,084 Bluesky posts, totaling 115,463 for embedding-based analysis.

US Congressional Tweets Dataset (2008–2017). To examine temporal shifts in political discourse and validate our bias detection models on a longitudinal corpus with known ground-truth affiliations, we incorporate the US Congressional Tweets Dataset [21], comprising tweets from verified Congressional member accounts spanning 2008 through 2017. This dataset captures a critical decade of political communication during which social media evolved from an experimental platform to a primary channel for political messaging. We apply our trained DistilBERT bias classifier to label all tweets in this corpus, enabling us to track the evolution of stance distributions over time within elite political discourse.

Politics.com Informal Political Discourse Corpus (2005). To evaluate out-of-domain and temporal generalization, we incorporate the Politics.com corpus [15], comprising 77,854 forum posts from 2005 with self-described political affiliations. This predates modern social media, capturing threaded discussions rather than microblogging. The corpus spans democrat (19,257), libertarian (9,672), conservative (9,042), independent (6,060), republican (4,319), liberal (3,909), and smaller groups, with 22,232 posts of unknown affiliation. We apply our DistilBERT bias classifier trained on Truth Social/Bluesky to predict Left/Neutral/Right labels and compare against self-identified affiliations. This tests whether contemporary social media models generalize to earlier forum discourse and whether partisan linguistic markers remained stable over two decades or evolved with platform migrations. Combined with our 2008–2017 Congressional analysis, this extends our longitudinal perspective to the pre-Twitter era, assessing whether ideological language is platform-invariant or shaped by venue-specific affordances.

4 Methods

Our goal is to explore practical and interpretable approaches for detecting political bias in short, informal social-media posts. Rather than committing to a single architecture, we compare a simple linear baseline with a compact transformer model and then stress-test the latter under domain and temporal shift. Specifically, we (i) establish a transparent linear baseline using Complement Naive Bayes, (ii) fine-tune a DistilBERT classifier on our Truth Social + Bluesky corpus, (iii) evaluate both models in-domain on that corpus, and (iv) apply the DistilBERT classifier to temporally and topically out-of-domain corpora (US Congressional tweets and Politics.com) to diagnose robustness and systematic biases.

4.1 Labeling Pipeline

Our labeling pipeline employs large language model assisted annotation to generate political bias labels at scale. We use Google's Gemini 2.5 Flash through the OpenRouter API, processing each post independently for Left, Right, or Neutral classification based on the stance expressed in its text.

The annotation prompt consists of two components. A system prompt frames the model as an expert political analyst tasked with classifying stance on a given topic. The user prompt then provides the post text and explicit definitions: Left indicates progressive positions, Right indicates conservative positions, and Neutral indicates factual or balanced content without clear partisan framing. To ensure deterministic outputs, we set temperature to zero and limit responses to ten tokens, forcing concise single-word classifications.

Post-processing normalizes the model output and applies fallback heuristics to handle edge cases, flagging ambiguous responses for manual review. The pipeline processes posts in batches with automatic checkpointing every twenty-five instances and introduces one-second delays between API calls for stable throughput. Each labeled post retains both the LLM-generated label and original metadata, enabling downstream quality analysis.

To validate the pipeline, we manually inspect stratified samples and compare LLM labels with annotator judgments, providing a coarse estimate of label noise. This approach balances scalability with a minimal level of quality control, but it does not replace full-scale human annotation; we leave large, systematically annotated validation sets to future work.

For subsequent experiments we distinguish two label regimes. We refer to the LLM-generated labels on the combined Truth Social and Bluesky corpus as the *gold* label set (while recognizing that they are machine-generated rather than human gold standard). We also construct a *noisy* label set that assigns platform-level labels only, treating all Bluesky posts as Left and all Truth Social posts as Right. This platform-as-label heuristic is not used to train our main models, but serves as an intentionally coarse alternative for ablation and baseline analysis.

4.2 Baselines

Platform-only prior. As a content-free validation baseline, we evaluate a rule-based classifier that uses only platform identity: on the Left/Right-only subset of the data it predicts Left for all Bluesky posts and Right for all Truth Social posts. This baseline provides an upper bound on what can be achieved from platform identity alone, and allows us to quantify how much additional signal our text models recover beyond platform-level information.

Naive Bayes (linear baseline). As a classical and interpretable baseline, we train a Complement Naive Bayes classifier on TF–IDF representations of posts using word unigrams and bigrams. Despite its strong independence assumptions, Naive Bayes remains effective for short text classification and provides a transparent, computationally efficient reference point with well-understood probabilistic semantics.

DistilBERT (main classifier). As our primary model, we fine-tune `distilbert-base-uncased` under the three label regimes described in Section 4.3: noisy binary (platform-as-label), gold binary (Left/Right only), and gold 3-class (Left/Neutral/Right). Posts are tokenized with a maximum length of 256 tokens, and we apply inverse-frequency class weights to mitigate label imbalance. We use a learning rate of 2×10^{-5}, batch size 32, weight decay 0.01, and fine-tune for three epochs on the combined Truth Social + Bluesky corpus. This compact encoder (66M parameters) serves as the main model for both in-domain evaluation and all cross-domain / temporal-shift experiments.

4.3 Training and Evaluation Setups

All experiments share the same stratified train/validation/test split of 131,808 posts (93,014 train, 13,240 validation, 25,554 test). For each post we store two labels:

Gold label (`gold_label`): the LLM-assisted Left/Neutral/Right label described in Section 4.1.

Noisy label (`noisy_label`): a platform-prior label that maps all Bluesky posts to *Left* and all Truth Social posts to *Right*.

The test set contains 10,223 Neutral, 8,997 Left, and 6,334 Right gold labels. We evaluate three supervised setups (and one non-learning baseline):

Noisy Binary. Models are trained on the `noisy_label` (platform-as-label) for all training instances, but evaluated on the subset of the test set where the gold label is Left or Right (15,331 posts). This setup answers: *How predictive is platform identity (and platform-correlated text) of Left/Right stance?*

Gold Binary. Models are trained and evaluated only on instances whose gold label is Left or Right. This provides the fairest text-only baseline for binary bias classification.

Gold 3-Class. Models are trained on the full training set using gold Left/Neutral/Right labels and evaluated on the full test set. This setup measures performance when Neutral is treated as an explicit abstention class.

Platform Rule (non-learning baseline). For comparison we define a trivial classifier that predicts *Left* for every Bluesky post and *Right* for every Truth Social post. We evaluate this rule on the same Left/Right test subset as the binary models.

We evaluate models using stratified train/validation/test splits on the combined Truth Social and Bluesky corpus. Our primary in-domain metrics are accuracy and Macro F1, supplemented by per-class precision and recall and confusion matrices. For temporal and out-of-domain analysis, we apply the DistilBERT classifier to US Congressional tweets (2008—2017) and the Politics.com forum (2005), using party labels and self-described affiliations as weak supervision signals as described below.

4.4 Cross-Domain Deployment

To assess temporal and out-of-domain generalization, we apply our DistilBERT bias classifier to two additional corpora: US Congressional tweets (2008–2017) and the 2005 Politics.com forum. For Congressional tweets, we use the verified member accounts provided by Yez-Feijo et al. [21] and treat party affiliation as a weak ground-truth signal: tweets from Democratic members are expected to be Left or Neutral, while tweets from Republican members are expected to be Right or Neutral. For Politics.com, we rely on the self-described political affiliations (*polafil*) provided in the corpus [15] and map them to coarse Left/Neutral/Right stance categories for diagnostic evaluation.

In both settings, we emphasize that these labels are *weak* and not equivalent to post-level ideological stance. Our goal is not to recover true ideology, but to characterize how a classifier trained on contemporary Truth Social and Bluesky discourse behaves under temporal and platform shift. We therefore analyze confusion matrices, row-normalized predicted distributions conditioned on affiliation, and aggregate statistics such as $P(\text{predicted stance} \mid \text{party})$ and $P(\text{predicted stance} \mid \text{polafil})$ to surface systematic biases and failure modes.

5 Results and Analysis

5.1 Exploratory Data Analysis

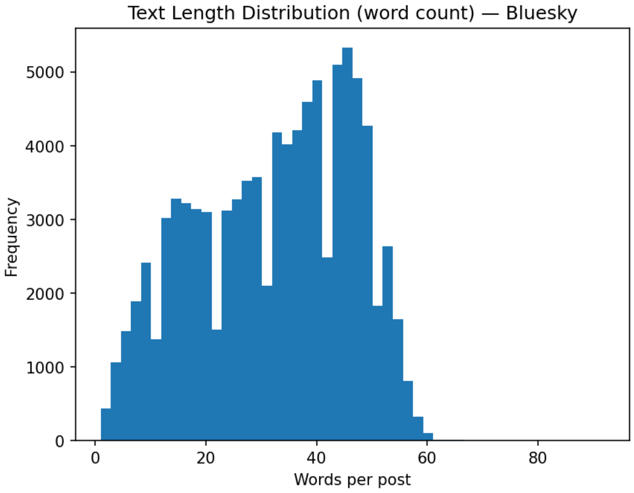

Figure 1: Bluesky post-length distribution.

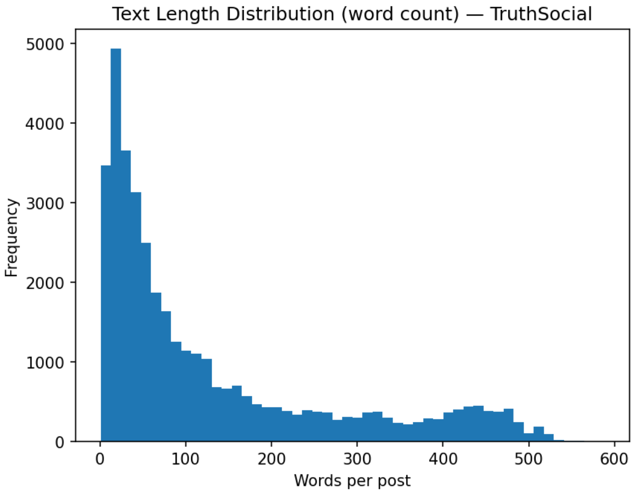

Figure 2: Truth Social post-length distribution.

Across platforms, post length reflects both technical constraints and community usage. Bluesky posts cluster tightly around medium lengths (roughly 40–60 words; Figure 1), consistent with its 300-character limit and a norm of compact commentary. Truth Social posts (Figure 2) show a sharp peak at very short lengths (10–20 words) plus a long tail of extended messages, reflecting frequent short reactions alongside occasional long-form rants. This variability means a classifier trained jointly on both platforms must cope with both extremely short, context-poor posts and much longer, more discursive content.

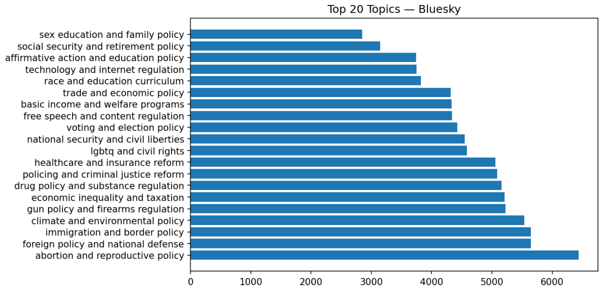

Figure 3: Bluesky topic distribution.

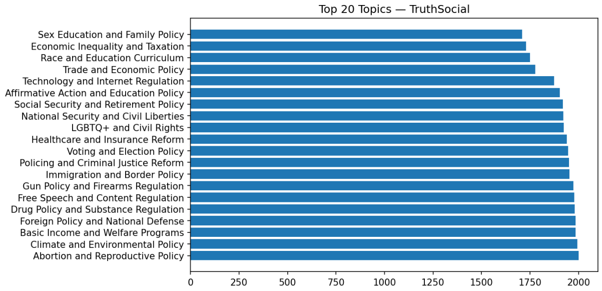

Figure 4: Truth Social topic distribution.

Topic coverage is broadly similar across platforms (Figures 3–4), with abortion, climate, immigration, and foreign policy among the most frequently discussed issues. However, Truth Social devotes relatively more volume to free-speech and content-regulation debates, taxation and economic grievance, and education/curriculum controversies, whereas Bluesky shows comparatively more content on healthcare, LGBTQ+ rights, and criminal-justice topics. These differences motivate our decision to stratify evaluation by topic and to interpret platform-level results through the lens of issue salience and framing.

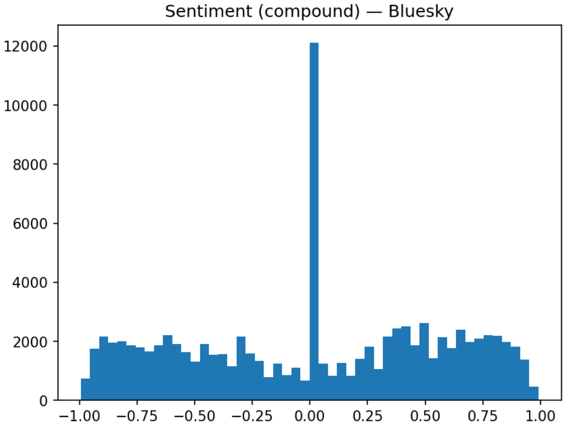

Figure 5: Bluesky sentiment histogram.

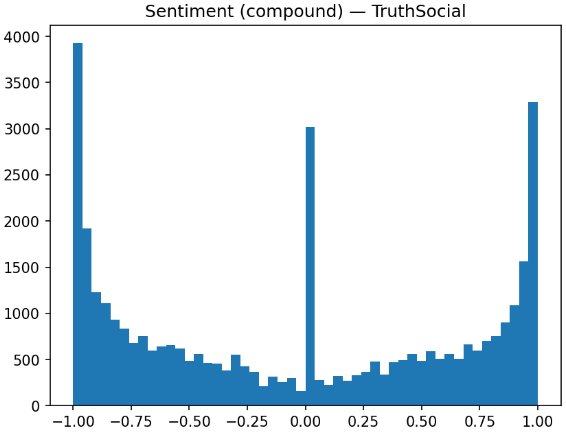

Figure 6: Truth Social sentiment histogram.

Sentiment profiles also diverge. Bluesky sentiment, measured with the VADER compound score, is dominated by a central mass near zero with a mild positive skew (Figure 5), consistent with a mix of informational and light social content. Truth Social shows a much more U-shaped distribution (Figure 6), with many strongly positive and strongly negative posts and relatively few neutral ones, suggesting higher emotional polarization.

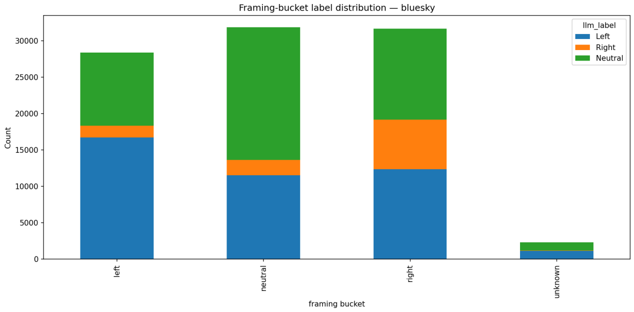

Figure 7: Bluesky label distribution by keyword framing bucket.

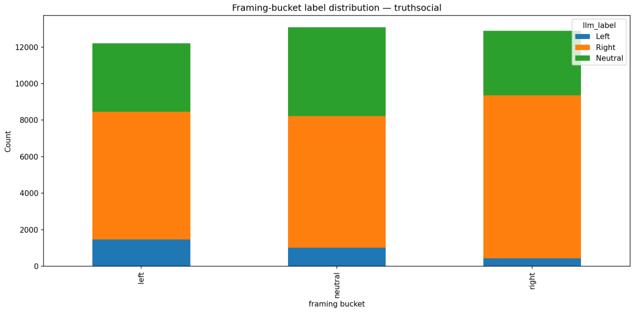

Figure 8: Truth Social label distribution by keyword framing bucket.

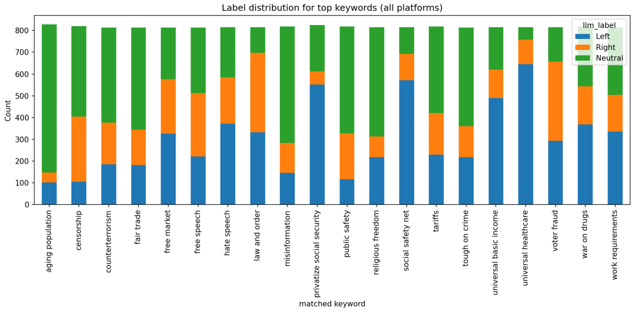

Figure 9: Label distribution for selected high-frequency keywords.

Finally, we sanity-check the LLM-generated labels by relating them to keyword-based framing buckets. On Bluesky, left-framed queries (e.g., *pro-choice*) yield predominantly Left labels, right-framed queries (e.g., *pro-life*) increase the share of Right labels, and neutral queries (e.g., policy terms) are mostly Neutral (Figure 7). On Truth Social, by contrast, Right labels dominate regardless of framing bucket (Figure 8), reflecting the platform's ideological homogeneity. At the keyword level (Figure 9), we observe three groups: descriptive policy terms that are mostly Neutral, one-sided cues that are strongly Left or Right, and contested rhetorical phrases (e.g., “free speech”) that attract substantial mass from both sides. These patterns suggest that framing-based keyword signals carry useful information on diverse platforms like Bluesky but must be interpreted cautiously on ideologically homogeneous ecosystems such as Truth Social.

5.2 Model Performance

Table 1: Naive Bayes classification report on the full test set (Gold 3-Class setup)

Class	Precision	Recall	F1-Score	Support
Left	0.592	0.765	0.667	8,997
Neutral	0.715	0.583	0.643	10,223
Right	0.671	0.593	0.630	6,334
Accuracy			0.650	25,554
Macro Avg	0.660	0.647	0.646	25,554
Weighted Avg	0.661	0.650	0.646	25,554

Naive Bayes. We establish baseline performance using Complement Naive Bayes on TF-IDF representations of 131,808 posts split into stratified train/validation/test sets (93,014 / 13,240 / 25,554). The corpus is distributed across Neutral (41%), Left (34%), and Right (26%) labels. We select alpha = 0.1 via three-fold cross-validation, achieving 65% accuracy and Macro F1 of 0.646 on the test set. Per-class F1 scores remain balanced at 0.667 (Left), 0.643 (Neutral), and 0.630 (Right) despite label imbalance. However, recall patterns reveal asymmetry: the model identifies Left posts most reliably (0.765) while struggling with Neutral (0.583) and Right (0.593), suggesting progressive discourse contains more distinctive lexical markers.

Error analysis reveals systematic confusion patterns. Right posts misclassify as Left nearly twice as often as Neutral (1,648 vs. 931), indicating partisan rhetoric—regardless of direction—shares stylistic features like emotional intensity that distinguish it from neutral content. Left posts confuse more with Neutral than Right (1,441 vs. 675), suggesting some progressive content adopts fact-based framing lacking overt markers. Most strikingly, Neutral posts are misclassified as Left (3,097) far more than Right (1,164), potentially indicating weak supervision bias or the model’s inability to distinguish factual reporting from left-leaning interpretation. These patterns motivate transformer experiments that may better capture subtle framing differences.

DistilBERT. We fine-tune DistilBERT-base-uncased on the same stratified splits (93,014 / 13,240 / 25,554), tokenizing posts to 256 tokens maximum and applying inverse-frequency class weights to address imbalance. Training proceeds for three epochs with learning rate 2e-5, batch size 32, and weight decay 0.01, totaling 8,721 gradient steps. Loss drops smoothly from 0.96 to 0.64 in epoch one, 0.50 in epoch two, and stabilizes around 0.40 in epoch three. Validation (75.4% accuracy, 0.751 Macro F1) and test (75.0% accuracy, 0.748 Macro F1) performance remain tightly aligned, confirming minimal overfitting despite 66M parameters.

Table 2: DistilBERT performance on validation and test sets

Metric	Validation	Test
Accuracy	0.754	0.750
Macro F1	0.751	0.748
Per-Class F1 Scores		
Left	0.748	0.756
Neutral	0.771	0.756
Right	0.734	0.731
Loss	0.640	0.643

The transformer substantially outperforms Naive Bayes (~10 percentage point gains) with far better class balance. Per-class F1 scores are nearly uniform: Left (0.756), Neutral (0.756), and Right (0.731). Unlike Naive Bayes, which showed high Left recall but struggled with Neutral (0.583) and Right (0.593), DistilBERT distributes errors evenly. The improvement for Neutral (0.643 to 0.756) is particularly striking, indicating contextualized representations better distinguish factual reporting from partisan framing. The near-parity between Left and Neutral F1 confirms that transformers reduce the asymmetric confusion that plagued the classical baseline, where Neutral posts were misclassified as Left (3,097 instances) far more than Right (1,164). This balanced performance provides a more reliable foundation for cross-platform generalization experiments.

Table 3: Binary Left/Right performance on the gold L/R test subset (15,331 posts).

Model / Setup	Accuracy	Macro F1
Platform rule	0.838	0.825
Naive Bayes (Noisy L/R)	0.764	0.737
Naive Bayes (Gold L/R)	0.812	0.798
DistilBERT (Noisy L/R)	0.812	0.796
DistilBERT (Gold L/R)	0.858	0.855

Platform-only baseline and binary L/R experiments. Table 3 summarizes accuracy and Macro F1 for the binary Left/Right setups on the gold-labeled test subset (15,331 posts). The platform-only rule already reaches 83.8% accuracy and 0.825 Macro F1. Naive Bayes trained on noisy platform-as-label supervision trails both this rule and the Naive Bayes model trained on gold L/R labels, while DistilBERT trained on gold L/R labels outperforms all baselines, including the platform rule.

Platform-only prior. The strong performance of platform-only prior is partly an artifact of the evaluation setup: after dropping all Neutral-labeled posts from the test set, the remaining distribution is highly skewed, with 81% of the surviving Bluesky posts labeled Left by the LLM and 90% of the surviving Truth Social posts labeled Right. In other words, on this filtered subset, platform ≈ stance:

$$P(\text{Left} \mid \text{Bluesky}, \text{L/R}) \approx 0.81$$
$$P(\text{Right} \mid \text{Truth}, \text{L/R}) \approx 0.90$$

This makes platform identity a very strong baseline for L/R classification.

Naive Bayes vs. platform. On the same L/R-only data, Complement Naive Bayes trained on the *gold* LLM labels reaches 81.2% accuracy and 0.798 macro F1, slightly below the platform-only baseline but above a Naive Bayes model trained on the *noisy* platform-as-label regime (76.4% / 0.737). Adding Neutral to form a three-way task depresses performance substantially: the 3-class Naive Bayes model yields 65.0% accuracy and 0.646 macro F1. Across all Naive Bayes setups, recall for Left is consistently higher (0.91–0.95) than for Right (0.53–0.69), indicating many Right to Left confusions and more distinctive lexical cues for left-coded content.

BERT vs. platform. For DistilBERT, the ranking is reversed: the *gold* binary L/R model outperforms the platform prior. DistilBERT trained on gold L/R labels reaches 85.8% accuracy and 0.855 macro F1, compared to 81.2% / 0.796 for DistilBERT trained on the *noisy* platform-as-label regime and 83.8% / 0.825 for the platform-only rule. As with Naive Bayes, introducing a Neutral class makes the task harder: the 3-class DistilBERT model attains 75.0% accuracy and 0.748 macro F1 (Left/Neutral/Right F1: 0.756/0.756/0.731). Validation and test scores remain closely aligned across all configurations, suggesting no significant overfitting.

Overall, these experiments yield three main insights. First, on the L/R-only subset, platform identity by itself is an extremely strong predictor of stance, and any content-based model should be compared against this baseline. Second, training on *gold* L/R labels consistently outperforms training directly on noisy platform-as-label supervision, both for Naive Bayes and DistilBERT, which supports the utility of the LLM-annotated corpus. Third, expanding from binary L/R to a three-way Left/Neutral/Right task substantially increases ambiguity and lowers headline metrics, but is necessary when one cares about neutral or abstaining content rather than purely polarized speech.

6 Cross-Domain and Temporal Shift Experiments

6.1 Congressional Tweets (2008–2017)

Applying our DistilBERT classifier to the US Congressional Tweets Dataset reveals a clear shift in how the model interprets elite political communication over time (Table 4). Under this classifier, predicted Neutral content declines steadily from 80.4% in 2008 to 48.3% in 2017, while both Left and Right labels increase. Right predictions more than double (15.2% to 34.4%), and Left predictions nearly quadruple (4.5% to 17.3%). The steepest decline in Neutral and growth in partisan labels occurs between 2015 and 2017, coinciding with the 2016 presidential election cycle.

We emphasize that these are *model-centric* trends: they show how a classifier trained on modern partisan platforms perceives congressional discourse, not a direct measurement of ground-truth ideology. In Section 6.2 we show that the same model exhibits substantial calibration bias on other corpora, suggesting that some portion of the apparent rightward lean reflects model behavior rather than purely substantive ideological change.

Table 4: Temporal Evolution of Stance Labels in Congressional Tweets (2008–2017)

Year	Left (%)	Neutral (%)	Right (%)
2008	4.5	80.4	15.2
2009	7.3	73.6	19.0
2010	8.9	66.9	24.1
2011	11.8	63.4	24.8
2012	11.1	64.1	24.8
2013	11.7	60.6	27.7
2014	11.8	61.2	27.0
2015	11.9	59.1	29.0
2016	10.8	56.5	32.7
2017	17.3	48.3	34.4

6.2 Politics.com Out-of-Domain Evaluation

To probe out-of-domain behavior, we evaluate the same classifier on the Politics.com forum corpus from 2005 [15]. Each post is associated with a self-declared political affiliation (*polafil*), including *democrat*, *republican*, *liberal*, *conservative*, *centrist*, and *fringe* categories. We apply the classifier to predict Left/Neutral/Right stance and examine the row-normalized distributions $P(\text{predicted stance} \mid \text{polafil})$:

Table 5: Row-normalized predicted stance distribution $P(\textbf{pred} \mid \textbf{polafil})$ on Politics.com, sorted by Pred Left, then Pred Neutral, Pred Right (all descending).

polafil	Pred Left	Pred Neutral	Pred Right
green	0.239	0.368	0.393
liberal	0.172	0.462	0.366
l-fringe	0.163	0.498	0.338
independent	0.129	0.434	0.437
democrat	0.126	0.479	0.394
centrist	0.125	0.357	0.518
libertarian	0.120	0.546	0.333
republican	0.107	0.506	0.387
conservative	0.099	0.416	0.485
r-fringe	0.075	0.375	0.550

The model recovers a coarse left–right ordering: explicitly right-coded affiliations such as *conservative*, *republican*, and *r-fringe* have the highest fraction of Right predictions, while *liberal*, *l-fringe*, and *green* have higher Left fractions. However, the absolute calibration is poor. For every affiliation group, the model predicts Right more often than Left, including for *democrat*, *liberal*, and *green*. Neutral predictions are also substantial across all groups.

When we collapse to a two-class setting and evaluate only on the ‘democrat’ and ‘republican’ subsets (19,257 and 4,319 posts respectively), treating Democrats as gold Left and Republicans as gold Right, overall accuracy drops to 17.4% and Macro F1 to 0.16.

It is important to note that these figures reflect a post-level evaluation against a user-level affiliation signal. A self-identified conservative user can still write an occasional post that expresses a liberal-leaning stance (e.g., praising a Democratic policy), and likewise a self-identified democrat can produce posts that endorse traditionally conservative positions. Since our classifier operates at the granularity of individual posts while *polafil* reflects a stable self-description of the author, disagreements between predicted stance and affiliation do not always indicate a classification error. Some portion of the low D vs. R accuracy therefore reflects genuine within-user variation in stance rather than purely model failure.

Nonetheless, the extremely low overall accuracy (17.4%) and the systematic underuse of the Left label across all affiliation groups suggest that label-mapping mismatch and classifier bias also play a substantial role, and that *polafil* should be treated as a weak supervision signal rather than gold-standard ground truth.

6.3 Error Analysis and Model Bias

To better understand the rightward skew observed on Congressional and Politics.com data, we conduct a targeted error analysis on Congressional tweets using party affiliation as a weak label. Restricting to Democratic members, we compare tweets the classifier predicts as Right vs. those it predicts as Left or Neutral. Structural features differ markedly: Democratic tweets predicted as Right are retweets 32% of the time, compared to 18% for those predicted as Left/Neutral, and they contain @-mentions in 66% of cases versus 48% for Left/Neutral. These patterns mirror structural cues in our training corpus, where right-coded posts on Truth Social and Bluesky heavily leverage retweets, mentions, and quoted content.

Lexical analysis with a Dem-only TF–IDF + logistic regression model further reveals that tokens most associated with Democratic tweets predicted as Right include *disarmhate*, *hillaryclinton*, *realdonaldtrump*, *housedemocrats*, *obama*, *obamacare*, and *muslimban*, whereas tokens most associated with Democratic tweets predicted as Left/Neutral include *gop*, *republicans*, *healthcare*, *medicare*, *climatechange*, and *lgbt*. In our modern training data, right-coded posts frequently mention Democratic leaders and causes in a critical way, while left-coded posts frequently mention Republican actors and policies. The classifier appears to transfer these associations to Congressional tweets, predicting Right whenever Democrats affirm Democratic figures or campaigns and predicting Left/Neutral when Democrats criticize Republicans.

Combined, these analyses suggest that the apparent rightward lean in our Congressional results is driven not solely by substantive ideology, but also by (i) structural cues such as retweet patterns and mentions, and (ii) lexical associations learned from contemporary partisan platforms where references to Democratic actors often occur in right-wing criticism. This reinforces the need to interpret temporal and cross-domain results as reflections of model biases and training-data artifacts rather than definitive measurements of ideological position.

7 Semantic Clustering Analysis

7.1 Text Embeddings and Clustering Algorithm

Dataset Overview. We constructed our clustering dataset by randomly sampling 80,000 posts from each of two publicly available platform corpora introduced in prior work [7, 8]. After removing reposts and discarding posts shorter than ten characters, the combined dataset analyzed in this study consists of 115,463 posts: 51,379 from Truth Social and 64,084 from Bluesky. These posts span diverse content domains, linguistic styles, and community norms, offering a unique opportunity to examine semantic overlap and divergence across ideologically and culturally distinct online environments. Standard preprocessing procedures were applied uniformly across both platforms. Posts containing only URLs or emojis were removed due to their lack of substantive textual content, while mixed posts containing meaningful text alongside media, links, or emojis were retained. The final dataset preserved sufficient semantic material for robust embedding and clustering, enabling a unified analysis of discourse patterns across both platforms.

Text Embeddings. To embed the posts into a shared semantic space, we employed the SentenceTransformer model `all-MiniLM-L6-v2`, which projects each post into a 384-dimensional vector representation. By encoding content from both platforms within the same embedding model, semantically related posts—regardless of origin—are mapped to neighboring regions in the high-dimensional space. This unified embedding space enables direct comparison of linguistic tendencies, topical structure, and stylistic variation across Truth Social and Bluesky.

To visualize the high-dimensional embeddings, we projected them into two dimensions using both Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). PCA provides a linear projection that emphasizes global variance structure, whereas UMAP performs non-linear manifold learning and is designed to preserve local neighborhood relationships. Together, these complementary methods offer a more complete view of the semantic landscape.

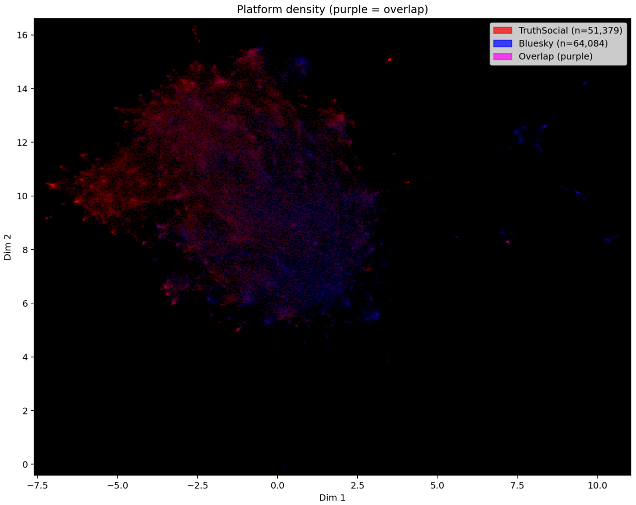

Figure 10: UMAP density visualization of the joint embedding space.

The UMAP density plot reveals a substantial region of purple overlap in the central area, indicating that a large portion of discourse from both platforms occupies a shared semantic core, with similar topics or linguistic patterns. At the same time, Truth Social posts (red) form denser concentrations along the upper-left region of the projection, while Bluesky posts (blue) dominate the right-hand and more peripheral areas. This pattern suggests that, although the two platforms share a common baseline of everyday conversation and general topics, each also develops distinct niche communities and topic clusters. The presence of isolated blue fragments on the far right of the UMAP projection corresponds to small, Bluesky-specific clusters that rarely appear on Truth Social, such as highly creative or multilingual content.

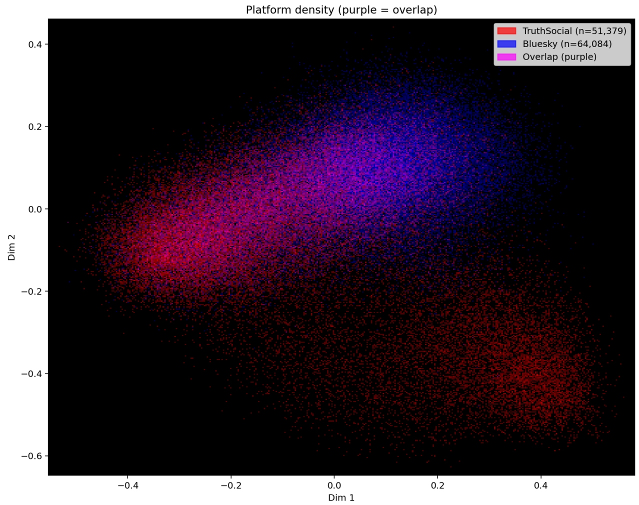

Figure 11: PCA density visualization of the joint embedding space.

The PCA density plot exhibits a broadly overlapping red and blue distribution, reinforcing the existence of shared semantic space. However, the second principal component differentiates the platforms: Truth Social content tends to skew toward lower PC2 values, while Bluesky content skews upward along PC2. This systematic separation implies consistent stylistic or topical differences between the platforms, even when they are embedded in a common representation space. As in the UMAP visualization, a purple core in the middle of the PCA projection reflects a shared semantic baseline, which likely corresponds to common conversational patterns, non-political social interactions, and cross-platform topics that are expressed similarly on both sites.

Clustering Algorithm and K Selection. We performed K-Means clustering on the embedding vectors to identify recurrent semantic themes across both platforms. Rather than selecting the number of clusters a priori, we evaluated a range of candidate values from 5 to 30 using three complementary cluster validity metrics: the Silhouette Score, the Calinski–Harabasz index, and the inertia-based Elbow method. To address computational constraints inherent in Silhouette Score computation for large datasets, a stratified sample of 5,000 posts was used while preserving relative representation from each platform. Elbow point detection was supplemented with second-derivative analysis to reduce subjectivity in identifying inflection points. Across all metrics, $k = 18$ emerged as the most stable and interpretable solution, achieving high separation between clusters, balanced distribution of posts, and coherent topical boundaries. The resulting clusters ranged from approximately 1,500 to over 9,000 posts, offering sufficient granularity to capture platform-specific themes while avoiding fragmentation of semantically similar content. A more detailed clustering summary is provided in Appendix C.

7.2 Cross-Platform Cluster Analysis

The 18 resulting clusters were grouped into five overarching domains (*Political/Ideological*, *Institutional/Public Affairs*, *Social/Cultural*, *Creative/Lifestyle*, and *Multilingual/Global Discourse*) to characterize the content ecosystems. The analysis confirmed pronounced thematic specialization between the two environments:

Truth Social. Truth Social content heavily concentrates in the *Political and Ideological* Domain, with clusters dominated by anti-Biden discourse, U.S. elections and partisan conflict, and patriotic or religious devotion. This emphasis extends into the *Institutional and Public Affairs* Domain, highlighting institutional skepticism (e.g., law enforcement, FBI, court-related discussions) and vaccine skepticism.

Bluesky. Bluesky content predominantly occupies the *Creative, Lifestyle, and General Interaction* Domain, including clusters focused on art and design sharing, gaming, food culture, and everyday conversation. Notably, Cluster 15, which captures multilingual interactions (Spanish, German, Persian, etc.), is composed of 96% Bluesky content, underscoring the platform's global reach and linguistic diversity—features almost entirely absent on Truth Social.

8 Key Findings and Platform Roles

Overall, the combined embedding and clustering analysis demonstrates that Truth Social and Bluesky operate as parallel yet distinct discourse ecosystems.

Truth Social. The platform functions primarily as a politically oriented, values-driven community space characterized by strong identity-based engagement.

Bluesky. In contrast, Bluesky serves as a broadly creative and socially interactive environment marked by greater linguistic diversity.

Despite this ideological and topical divergence, several clusters reveal cross-platform convergence on themes such as general conversation, humor, moral reflection, and economic concerns. These shared thematic regions underscore underlying human preoccupations that remain interconnected across platforms, even as each environment cultivates its own distinct community identity.

9 Discussion and Limitations

Our results highlight both the promise and the pitfalls of cross-platform bias detection. On our Truth Social and Bluesky corpus, LLM-derived labels and contextualized encoders yield strong in-domain performance, but a trivial platform-only heuristic already attains 83.8% accuracy on the L/R subset. Our best DistilBERT model trained on gold L/R labels exceeds this baseline, indicating that there is recoverable text-level signal beyond platform identity, yet platform remains an unusually strong confound that must be accounted for in evaluation. Additionally, the Congressional and

Politics.com experiments reveal substantial limitations when these models are applied under temporal and domain shift.

First, our labels are derived from a large language model pipeline rather than exhaustive manual annotation. Although spot checks suggest that the labels are usable at scale, they inevitably encode the biases of the underlying LLM and of our prompt and labeling design. Second, our primary training data come from two contemporary, English-language platforms with particular community compositions. When the resulting classifier is applied to Congressional tweets and a 2005 forum, we observe overprediction of Neutral and Right labels, underuse of the Left label, and poor alignment with self-declared political affiliations. These behaviors caution against treating model outputs as ground-truth ideology without careful calibration.

Third, our temporal analysis of Congressional tweets demonstrates how a modern social-media-trained classifier perceives increasing polarization over time, but we cannot disentangle model bias from true ideological change with the current labels. Future work could incorporate manual annotation or additional weak signals (e.g., roll-call votes, bill sponsorship) to obtain more reliable ground truth. Finally, our semantic clustering and topic analyses focus on English-language text and do not directly generalize to multilingual or non-text modalities.

Despite these limitations, we argue that such diagnostic experiments are valuable in their own right: they surface where bias classifiers fail, how platform-specific patterns bleed into cross-domain predictions, and what kinds of linguistic cues models rely on when classifying political language across time and media.

10 Future Work

Several extensions could strengthen this work. First, manual annotation of a stratified validation set would enable rigorous assessment of LLM label quality, calibration, and systematic error patterns. Second, incorporating additional weak signals such as Congressional roll-call votes or bill sponsorship could provide more reliable temporal ground truth and support better calibration of model scores. Third, extending the analysis beyond English and text-only posts—for example, to multilingual content and visual or multimodal political messaging—would test whether the identified patterns generalize across languages and modalities. Fourth, exploring domain-adaptive pretraining and stance-to-bias transfer (e.g., from SemEval stance datasets or PoliBERTweet-style encoders) could improve robustness under platform and temporal shifts. Finally, carefully monitored deployment of these classifiers in near real time to track evolving platform dynamics would reveal how quickly partisan discourse patterns shift in response to major political events, while also providing continuous feedback on model biases and failure modes.

Taken together, our results show that an LLM-labeled, cross-platform corpus combined with compact transformers can capture meaningful ideological patterns while also inheriting strong platform and model biases. We hope that the dataset, baselines, and diagnostic analyses introduced here provide a practical foundation for more robust and transparent studies of political language online.

References

[1] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. arXiv:2010.05338 [cs.CL] https://arxiv.org/abs/2010.05338

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of EMNLP*.

[3] Emily Chen, Ashok Deb, and Emilio Ferrara. 2021. Election2020: the first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science* 5, 1 (April 2021), 1–18. doi:10.1007/s42001-021-00117-9

[4] Ernesto Colacrai, Federico Cinus, Gianmarco De Francisci Morales, and Michele Starnini. 2024. Navigating Multidimensional Ideologies with Reddit's Political Compass: Economic Conflict and Social Affinity. arXiv:2401.13656 [cs.SI] https://arxiv.org/abs/2401.13656

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[6] Tiziano Fagni and Stefano Cresci. 2022. Fine-grained prediction of political leaning on social media with unsupervised deep learning. *Journal of Artificial Intelligence Research* 73 (2022), 633–672.

[7] Andrea Failla and Giulio Rossetti. 2024. "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PLOS ONE* 19, 11 (Nov. 2024), e0310330. doi:10.1371/journal.pone.0310330

[8] Patrick Gerard, Nicholas Botzer, and Tim Weninger. 2023. Truth Social Dataset. arXiv:2303.11240 [cs.SI] https://arxiv.org/abs/2303.11240

[9] Julie Jiang, Xiang Ren, and Emilio Ferrara. 2023. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks. arXiv:2207.08349 [cs.SI] https://arxiv.org/abs/2207.08349

[10] Kornraphop Kawintiranon and Lisa Singh. 2022. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. 7360–7367.

[11] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, et al. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

[12] Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech Language* 63 (2020), 101075. doi:10.1016/j.csl.2020.101075

[13] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2355–2365.

[14] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[15] Tony Mullen and Robert Malouf. 2006. A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. https://api.semanticscholar.org/CorpusID:6570757

[16] Daniel Preotiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle H. Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:27843613

[17] Kamalakkannan Ravi and Adan Ernesto Vela. 2024. Comprehensive dataset of user-submitted articles with ideological and extreme bias from Reddit. *Data in Brief* 56 (2024), 110849. doi:10.1016/j.dib.2024.110849

[18] Peyman Rostami, Vahid Rahimzadeh, Ali Adibi, and Azadeh Shakery. 2025. PolitiSky24: U.S. Political Bluesky Dataset with User Stance Labels. arXiv:2506.07606 [cs.CL] https://arxiv.org/abs/2506.07606

[19] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *Comput. Surveys* 34, 1 (March 2002), 1–47. doi:10.1145/505282.505283

[20] Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. 2024. An evaluation of language models for hyperpartisan ideology detection in Persian Twitter. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*. 51–62.

[21] Oscar Yáñez-Feijóo. 2023. US Congressional Tweets Dataset. https://www.kaggle.com/datasets/oscaryezfeijo/us-congressional-tweets-dataset

[22] Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 371–378.

Appendix A: Framings and Keywords

To construct the topic-annotated corpus and to probe how framing correlates with LLM-assigned bias labels, we curated a set of topic-specific search templates. For each political issue, we manually defined three keyword buckets: *Left-framed*, *Right-framed*, and *Neutral* terms. These were used to seed platform search APIs and to assign each retrieved post to a *framing bucket* based on the query that surfaced it. As discussed in the main text, these keywords were *not* used as supervision signals for the classifier; instead, they serve two purposes: (i) ensuring coverage across a wide range of politically salient topics, and (ii) enabling post-hoc sanity checks on the LLM labels by comparing label distributions across left/right/ neutral query framings.

Table 6 summarizes the topics and example search phrases used in each framing bucket.

Table 6: Political topics and example search queries by framing bucket.

Topic	Left-framed keywords	Right-framed keywords	Neutral keywords
Abortion and Reproductive Policy	reproductive rights; pro-choice; abortion access; women's rights	pro-life; unborn; sanctity of life; heartbeat bill	abortion law; abortion policy; Roe v. Wade; Planned Parenthood
Gun Policy and Firearms Regulation	gun control; assault weapons ban; background checks	gun rights; Second Amendment; 2A; constitutional carry	firearms policy; gun laws; gun ownership
Climate and Environmental Policy	climate crisis; renewable energy; Green New Deal	energy independence; fossil fuels; climate hoax	climate change; carbon emissions; environmental regulation; Paris Agreement
Healthcare and Insurance Reform	universal healthcare; Medicare for All; public option	health freedom; government overreach; private insurance	healthcare reform; Affordable Care Act; health insurance policy
Immigration and Border Policy	immigrant rights; DACA; asylum seekers; family separation	border crisis; illegal immigration; build the wall	immigration policy; border security; visa policy
LGBTQ+ and Civil Rights	LGBTQ+ rights; trans rights; marriage equality	religious freedom; parental rights; traditional values	civil rights; anti-discrimination; gender identity policy
Voting and Election Policy	voting rights; voter suppression; expand mail-in voting	election integrity; voter fraud; secure elections	election reform; voter ID laws; ballot access
Economic Inequality and Taxation	wealth tax; economic justice; raise minimum wage	tax burden; job creators; free market	tax policy; income inequality; economic mobility
Policing and Criminal Justice Reform	police reform; defund the police; mass incarceration	law and order; back the blue; crime wave	criminal justice reform; public safety; police accountability
Free Speech and Content Regulation	hate speech; content moderation; misinformation	free speech; censorship; cancel culture	First Amendment; online speech; platform policy
Affirmative Action and Education Policy	affirmative action; diversity in education; racial equity	merit-based admissions; colorblind policy; reverse discrimination	college admissions; education policy; Supreme Court decision
Drug Policy and Substance Regulation	drug decriminalization; harm reduction; marijuana legalization	war on drugs; fentanyl crisis; tough on crime	drug policy; opioid epidemic; controlled substances
Foreign Policy and National Defense	diplomacy; humanitarian aid; anti-war	military strength; America First; peace through strength	foreign policy; NATO; defense spending; Ukraine war
Technology and Internet Regulation	tech accountability; data privacy; AI regulation	free market innovation; anti-censorship; Big Tech bias	technology policy; Section 230; AI ethics
Race and Education Curriculum	racial justice education; anti-racism curriculum; diversity training	critical race theory; parental rights in education; woke indoctrination	education curriculum; teaching history; school policy
Sex Education and Family Policy	comprehensive sex education; LGBTQ-inclusive curriculum	abstinence education; parental consent; family values	sex education policy; school curriculum; health education
Basic Income and Welfare Programs	universal basic income; social safety net; poverty relief	welfare dependency; personal responsibility; work requirements	welfare policy; income support; economic assistance
Trade and Economic Policy	fair trade; labor rights; climate-friendly trade	free trade; tariffs; America First trade	trade policy; import/export; trade agreements
Social Security and Retirement Policy	protect social security; expand benefits	entitlement reform; reduce spending; privatize social security	retirement policy; social security funding; aging population
National Security and Civil Liberties	surveillance reform; privacy rights; anti-war movement	national security; border protection; patriotism	counterterrorism; cybersecurity; civil liberties

Appendix B: Labeling Prompts

For all Truth Social and Bluesky posts in our main corpus, we obtain LLM-assisted stance labels using Google's Gemini 2.5 Flash accessed via the OpenRouter API. Each post is labeled independently with one of three classes: *Left*, *Neutral*, or *Right*. The model receives a system message that frames it as an expert political analyst and a user message that provides the post text and the topic used for keyword retrieval.

The system message is:

```
You are an expert political analyst. Your task is to classify the
political stance of a social media post regarding the specific topic of: '{topic}'.
```

The user message template is:

```
Post text:
"{post_text}"

Based on the post text, classify its political stance on the topic of '{topic}'.
Respond with ONLY one of the following words:
- Left: The post expresses a clear left-leaning or progressive stance on the topic.
- Right: The post expresses a clear right-leaning or conservative stance on the topic.
- Neutral: The post is purely factual, balanced, or does not take a clear
  political stance on the topic.
```

We send these messages using the Chat Completions API with temperature set to 0.0 to encourage deterministic behavior and `max_tokens` set to 10 to enforce short outputs. The raw model response is post-processed by a small cleaning function that trims whitespace, lowercases the output, and maps any variant such as "left." or "LEFT" to one of the three allowed labels (`Left`, `Neutral`, `Right`). Responses that fall outside this set are rare and are either discarded or flagged for manual inspection in our labeling pipeline.

Appendix C: Cluster Summary

Table 7: Summary of cluster composition and topic focus for Bluesky and Truth Social posts

Cluster	Bluesky (%)	Truth Social (%)	Dominant	Topic Summary	Representative Terms
0	64.7	35.3	Bluesky	Platform migration and meta discussion about social media environments	people, bluesky, post, media, app
1	58.7	41.3	Mixed	Gender discourse, feminism, and women's issues	woman, girl, women, lady, love
2	74.4	25.6	Bluesky	Online content, tech, gaming, and digital art culture	com, www, game, twitch, app
3	11.4	88.6	Truth Social	Anti-Biden discourse and political critique of the administration	biden, joe biden, president, hunter
4	84.1	15.9	Bluesky	Movies, books, entertainment, and creative writing discussions	movie, book, watch, love, story
5	50.1	49.9	Balanced	Economic concerns, inflation, and social commentary	money, gas, tax, inflation
6	93.5	6.5	Bluesky	Artistic expression, drawings, commissions, and design sharing	art, drawing, design, sketch
7	24.4	75.6	Truth Social	Law enforcement, FBI, court discussions, and institutional trust	fbi, court, police, judge
8	82.7	17.3	Bluesky	Food culture, daily lifestyle content, and light humor	food, eat, coffee, pizza
9	79.2	20.8	Bluesky	General conversation, memes, jokes, and casual remarks	just, know, time, love, think
10	43.5	56.5	Truth Social	Anti-vaccine messaging and pandemic skepticism	covid, vaccine, pandemic, virus
11	52.2	47.8	Mixed	Social and moral debate including religion, rights, and gender issues	trans, god, children, rights
12	52.3	47.7	Bluesky	Pop culture, online discourse, and satire-driven commentary	guy, man, time, did, know
13	68.0	32.0	Bluesky	Holiday greetings, personal updates, and daily life reflections	day, today, happy, morning
14	19.0	81.0	Truth Social	U.S. elections, partisan narratives, and political mobilization	trump, election, democrats, gop
15	96.0	4.0	Bluesky	Multilingual content spanning Spanish, German, Persian, and others	que, en, die, und, ich
16	0.7	99.3	Truth Social	Religion, national identity, and patriotic devotion	god, love, trump, president
17	46.4	53.6	Truth Social	Geopolitics, global conflicts, and international crises	ukraine, russia, israel, war