Exploiting contextual information to improve stance detection in informal political discourse with LLMs*

Arman Engin Sucu[1], Yixiang Zhou[2], Mario A. Nascimento[2], Tony Mullen[1]
Khoury College of Computer Sciences
Northeastern University
[1]Seattle Campus, USA [2]Vancouver Campus, Canada
Correspondence: a.mullen@northeastern.edu

Abstract

This study investigates the use of Large Language Models (LLMs) for political stance detection in informal online discourse, where language is often sarcastic, ambiguous, and context-dependent. We explore whether providing contextual information, specifically user profile summaries derived from historical posts, can improve classification accuracy. Using a real-world political forum dataset, we generate structured profiles that summarize users' ideological leaning, recurring topics, and linguistic patterns. We evaluate seven state-of-the-art LLMs across baseline and context-enriched setups through a comprehensive cross-model evaluation. Our findings show that contextual prompts significantly boost accuracy, with improvements ranging from +17.5% to +38.5%, achieving up to 74% accuracy that surpasses previous approaches. We also analyze how profile size and post selection strategies affect performance, showing that strategically chosen political content yields better results than larger, randomly selected contexts. These findings underscore the value of incorporating user-level context to enhance LLM performance in nuanced political classification tasks.

1 Introduction

Political stance detection is an increasingly relevant part of analyzing the flow of ideas in online environments where discourse is informal and implicitly expressed. Understanding a text or individual's ideological standpoint can be helpful for applications such as content moderation, public opinion tracking, and misinformation detection. Approaches to political stance detection using traditional natural language processing (NLP) and machine learning methods have been closely related to approaches to sentiment analysis.

However, political language is often nuanced and tends to be comparable to relatively difficult sentiment analysis domains. Posts with political stance on social networks are often ambiguous, sarcastic, or context-dependent. For example, consider the statement: *"Great, another tax cut for the rich—just what we needed!"*. Without additional context, this could either express support or sarcasm. Political intent is often embedded in subtext or prior engagement, which traditional models fail to capture (Mullen and Malouf, 2006; Malouf and Mullen, 2008; Samih and Darwish, 2021).

While earlier methods such as lexicon-based classifiers or keyword matching approaches perform poorly on such nuanced input, recent advancements in LLMs such as GPT-4 (OpenAI, 2024), LLaMA (AI, 2024a), and DeepSeek (et al., 2025) offer promise in handling complex language understanding (Cao and Drinkall, 2024; Kim et al., 2024).

The emergence of LLMs has fundamentally transformed approaches to sentiment analysis and stance detection. Traditional methods based on lexicons, feature engineering, and specialized classifiers have been largely supplanted by these general-purpose models that can capture subtle linguistic nuances, contextual cues, and implicit sentiment without task-specific architectures (Cruickshank and Ng, 2024; Allaway and McKeown, 2023). However, despite this paradigm shift, the core challenge of contextual understanding remains (Bhattacharya et al., 2024).

Nonetheless, even state-of-the-art LLMs struggle with implicit political signals, ideological ambiguity, and sarcastic cues. Our project investigates whether political stance can be reliably classified by augmenting LLM predictions with contextual cues, building on previous research that demonstrated the value of contextual information in political classification tasks (Mullen and Malouf, 2006; Malouf and Mullen, 2008; Doddapaneni et al., 2024).

*Dataset: https://github.com/tonymullen/politics.com. Code: https://github.com/armanengin/contextual-stance-llms.

In this study, we introduce a contextual enrichment framework that supplements LLM input with user profile summaries derived from historical forum posts. These profiles include inferred political leaning, recurring discussion topics, and linguistic patterns (Wu et al., 2024; Ye et al., 2021). By providing this additional context, we aim to improve stance classification accuracy—especially for posts that are short, ambiguous, or stylistically neutral.

We evaluate this approach on a real-world political forum dataset, comparing baseline classification against context-enhanced setups through a comprehensive cross-model evaluation of seven state-of-the-art LLMs. Our results show that incorporating profile-level context significantly improves model performance, with absolute accuracy gains ranging from +24.5% to +38.5%. We further investigate how profile size and post selection strategies affect performance, revealing that strategically selected political content contributes more than sheer volume (Cao and Drinkall, 2024; Welch et al., 2022).

This work highlights the importance of integrating user-level context into prompt design for political NLP tasks and offers a scalable method for enhancing classification reliability in informal discourse settings.

2 Related Work

Political stance detection spans multiple research traditions, from early sentiment analysis to recent LLM-based approaches. We review work in three key areas: (1) political stance classification techniques, (2) contextual enrichment methods, and (3) personalization for language models.

2.1 Political Stance Classification

Political sentiment analysis has long informed efforts to identify ideological positions in text. Early work focused on classifying opinion polarity in political tweets or news, often using lexicons or shallow models (Mohammad et al., 2017; Caetano et al., 2018). Studies also highlighted the role of affect in political discourse and the asymmetry of negative sentiment spread (Antypas et al., 2023; Sen et al., 2020). More recent research developed domain-specific and multilingual models to better capture political meaning in social media content (Aquino et al., 2025; Kawintiranon and Singh, 2022).

Building on this foundation, political stance detection has progressed from rule-based and lexicon-driven methods to neural and prompt-based approaches. Early studies explored user-level classification in online forums using discourse features (Mullen and Malouf, 2006; Malouf and Mullen, 2008; Samih and Darwish, 2021; Zhou and Elejalde, 2024), highlighting challenges posed by implicit and informal political language. While these approaches laid important groundwork for modeling user-level political stance, they lacked the contextual understanding capabilities that our approach leverages.

2.2 Contextual LLM Approaches

Recent LLMs enable zero- and few-shot stance classification without task-specific models. Prompting strategies with metadata or topic cues improve accuracy (Cao and Drinkall, 2024; Cruickshank and Ng, 2024; Kim et al., 2024; Allaway and McKeown, 2023). User-level modeling further boosts performance by leveraging behavioral or linguistic summaries (Bhattacharya et al., 2024; Doddapaneni et al., 2024; Welch et al., 2022; Wu et al., 2024; Ye et al., 2021). Evaluations on social media platforms like Twitter/X demonstrate model potential and limitations (Gambini et al., 2024), while frameworks like DEEM dynamically adapt to user history (Wang et al., 2024). Our work extends these approaches by systematically exploring how different types of user-level context affect classification accuracy across diverse LLM architectures.

2.3 Personalization and Reasoning in LLMs

Personalization in LLMs has advanced through techniques such as persona-aware attention, guided profile generation, retrieval-augmented prompting, and adaptive calibration. These methods have shown strong performance across dialogue, writing assistance, and recommendation tasks (Huang et al., 2023; Zhang, 2024; Salemi et al., 2024; Tan et al., 2024; Mysore et al., 2024). Recent work also highlights the importance of preference alignment, with studies evaluating how well LLMs follow user-specific instructions in downstream tasks (Zhao et al., 2025).

Complementary to these personalization efforts, recent research has explored reasoning-aware prompting strategies—such as Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022), ReAct (Yao et al., 2023), AutoPrompt (Shin et al., 2020), and prefix-tuning (Li and Liang, 2021)—which aim to improve model understanding of implicit, ambiguous, or sarcastic cues. While

our approach does not employ these methods, they represent promising future directions. Techniques like Chain of Preference Optimization (Zhang et al., 2024), which integrate user preferences into multi-step reasoning, may further enhance stance detection when combined with contextual enrichment.

Our work focuses specifically on user-level contextual prompting. By enriching model input with structured user profiles, we show consistent improvements in stance classification across seven state-of-the-art LLMs. These findings highlight the value of user-informed prompting in capturing nuanced signals in political discourse, and they may complement reasoning-based approaches in future hybrid systems.

3 Dataset and Preprocessing

Our study utilizes a political discourse dataset originally compiled by Mullen and Malouf (2006), consisting of approximately 77,854 posts downloaded from discussions on politics.com. The dataset is organized into topic threads, chronologically ordered, and identified according to author and author's stated political affiliation.

3.1 Data Source and Characteristics

The dataset contains contributions from 408 unique users engaged in various political discussions. User posting activity follows an inverse power-law distribution typical of online communities, with 77 posters (19%) contributing only a single post. The most active user contributed 6,885 posts, followed by the second most active with 3,801 posts.

A key feature of this dataset is that users self-declared their political affiliations, providing ground truth labels for our classification task.

Figure 1 shows the distribution of political affiliations in the dataset, which is relatively balanced between major ideological groups.

RIGHT 34%	Republican	53
	Conservative	30
	R-fringe	5
LEFT 37%	Democrat	62
	Liberal	28
	L-fringe	6
OTHER 28%	Centrist	7
	Independent	33
	Libertarian	22
	Green	11
	Unknown	151

Figure 1: Distribution of posts in the data by general class and by a slightly modified version of the writers' own self-descriptions.

3.2 Data Preprocessing

For our experiments, we processed this dataset in several key ways:

1. We mapped the original fine-grained political affiliations into three broad categories: LEFT (Democrat, Liberal, Left-fringe), RIGHT (Republican, Conservative, Right-fringe), and UNKNOWN (all other labels including Centrist, Independent, Libertarian, and Green).

2. We focused only on users with clear LEFT or RIGHT labels, filtering out posts from users with UNKNOWN political affiliation. This resulted in a filtered dataset of 56,035 posts from 257 users with declared political leanings.

3. For each user with a known political affiliation, we split their posts into two sets: 70% for profile generation (used to create user context) and 30% for testing classification performance (reserved for evaluation). We used a fixed random seed (42) for this split to ensure reproducibility across experiments and enable direct comparison of results.

4. We maintained post structure and metadata throughout preprocessing by preserving quote markers to differentiate between original content and quoted text, keeping forum-specific formatting to maintain conversational context, and retaining chronological ordering within each user's posts.

This approach allowed us to maintain the informal, conversational nature of the discourse while creating a structured dataset suitable for both baseline and context-enriched classification experiments. To ensure experimental rigor, we used the same test set for all experiments, allowing direct comparison between baseline and context-enhanced approaches.

4 Methodology and Experimental Design

Our approach centers on how contextual information about users' past behaviors can enhance LLMs' ability to classify political stance in informal discourse. We conducted three distinct experiments to thoroughly investigate the effectiveness of contextual enrichment.

4.1 Experimental Framework Overview

4.1.1 Implementation Approach

We define this as a binary stance classification task. Each input consists of a single forum post authored by a user. In the baseline setup, the post is provided to the model in isolation. In the context-enriched setup, the same post is preceded by a structured user profile summarizing the author's historical political behavior. The model is prompted to return a JSON object containing a predicted stance label—LEFT or RIGHT—and an accompanying explanation. The ground truth label is derived from the user's self-declared political affiliation in the dataset.

All experiments shared a common implementation approach to ensure consistent results. We accessed the LLMs through a unified API interface, providing standardized access across different model architectures. To maintain consistency, we applied identical parameters across all experiments: temperature set to 0.1 to minimize stochastic variation, standardized JSON output format for automated evaluation, and identical prompt structures except for the addition of context. Throughout our experiments, we evaluated two classification pipelines: a baseline where models classify posts without any user context, and a context-enriched approach where the same posts are classified with user profiles prepended in the prompt.

4.1.2 Experimental Progression

We implemented three sequential experiments, with each building on findings from the previous:

1. **Contextual Enrichment Impact**: Evaluating the maximum potential benefit of user profiles for classification accuracy
2. **Context Optimization Framework**: Determining optimal post selection strategies and volume for profile generation
3. **Cross-Model Performance Analysis**: Assessing different LLMs' capabilities in both profile generation and classification roles

4.2 User Profile Structure

Across all experiments, we used a consistent structured format for user profiles. Each profile contained the inferred political stance (left, right, or unknown) based on consistent ideological signals, the model's self-assessed confidence in its stance assignment (high, medium, or low), 3–5 specific linguistic or topical indicators supporting the assigned leaning, a list of common subjects the user discusses, a qualitative summary of the user's tone, a description of whom the user supports or criticizes, and optional free-text insights. These fields were generated using a structured prompt (see Appendix A.1), emphasizing objectivity, pattern recognition, and valid JSON formatting.

4.3 Experiment 1: Contextual Enrichment Impact

Our first experiment aimed to establish whether user profiles could improve classification performance and to measure the maximum potential benefit. We used Gemini 2.0 Flash (DeepMind, 2024) (with its 1M token context window) to generate comprehensive user profiles from all available posts in the profile-building set. Unlike later experiments, we did not selectively sample posts but instead used all available posts per user to generate the most comprehensive profiles possible. We evaluated on a set of 200 reserved test posts, ensuring a balanced representation of different political orientations. This experiment established the ceiling performance for our contextual enrichment approach.

4.4 Experiment 2: Context Optimization Framework

After establishing the effectiveness of contextual enrichment, we investigated how to optimize the context generation process. We implemented and evaluated five distinct post selection strategies: The **PoliticalSignalSelection** strategy prioritizes posts with strong political content by using a weighted lexicon of political terms in three categories: general political terms (e.g., 'politics', 'government', 'vote') with weight 1, party-specific terms (e.g., 'democrat', 'republican', 'liberal') with weight 2, and hot-button issues (e.g., 'abortion', 'gun', 'immigration') with weight 3. It calculates a political signal score for each post based on term frequency, boosts scores for posts in political subforums (+5 points), adds small random noise (0–1) to break ties, and selects 60% highest-scoring posts and 40% diverse-topic posts (see Appendix B for full implementation details).

Prior studies have shown that content-based filtering—specifically targeting politically salient posts—significantly improves stance detection accuracy. Aldayel and Magdy (2019) and Preotiuc-Pietro et al. (2017) demonstrate that features derived from political lexicons and issue-related key-

words are more informative than post recency or length. Rahimzadeh et al. (2025) further showed that filtering timelines to remove off-topic content enhanced user profiling with LLMs in large-scale settings. These findings support our PoliticalSignalSelection strategy: by scoring and selecting posts with high ideological signal, we retain the most diagnostic content for modeling user stance.

We also tested **RandomSelection** (randomly samples posts without consideration for content), **ControversialTopicSelection** (prioritizes posts containing terms from contentious political topics using a library of 150+ controversial keywords), **RecentPostSelection** (selects the most recent posts from a user's history), and **LongFormSelection** (prioritizes longer posts based on word count).

We evaluated eight different post count settings to understand the relationship between context volume and classification performance, ranging from minimal context (1, 2, 3 posts), medium context (5, 10 posts), and extensive context (20, 30 posts), to maximum context (50 posts). We tested each combination of post count and selection strategy, resulting in 40 distinct experimental conditions (8 post counts × 5 selection strategies). Each condition was tested on up to 50 users with 5 test posts per user (max 250 classification instances per condition), for a total of approximately 10,000 classification instances across all conditions.

Through this experiment, we determined that **PoliticalSignalSelection** with 10-20 posts yielded near-optimal results, with diminishing returns beyond this threshold.

4.5 Experiment 3: Cross-Model Performance Analysis

Our final experiment investigated how different LLMs perform in both profile generation and classification roles, using the optimized parameters from Experiment 2. We tested seven state-of-the-art LLMs representing diverse architectures: Claude 3.7 Sonnet (Anthropic, 2024), Grok-2-1212B (xAI, 2024), GPT-4o Mini (OpenAI, 2024), Mistral Small-24B (AI, 2024b), Meta-LLaMA 3.1-70B (AI, 2024a), Qwen (Cloud, 2024), and Gemini 2.0 Flash (DeepMind, 2024).

Based on findings from Experiment 2, we standardized parameters across models, using only the PoliticalSignalSelection strategy, 50 posts per user profile, and the same test dataset of 200 posts per model. We implemented a 7×7 experimental design where each model generated user profiles for the same set of users, each model was then used to classify posts using profiles created by every model, and all 49 model combinations were evaluated using the same test dataset. This comprehensive evaluation revealed which models excel at generating informative profiles and which are most effective at leveraging contextual information for classification.

4.6 Evaluation Approach

To assess the impact of contextual enrichment across our experiments, we focused on several key comparative metrics. We measured absolute improvement as the percentage point difference between context-enriched and baseline accuracy, directly quantifying the benefit of providing user profiles. We analyzed the relative impact across models by examining how improvement correlates with baseline performance, revealing whether weaker models benefit more from contextual information. We studied context efficiency as performance relative to context volume, helping identify the optimal balance between context size and computational requirements. Finally, we analyzed cross-model complementarity, determining which model combinations (profile generator + classifier) yield the best performance and reveal potential complementary strengths.

5 Results and Analysis

5.1 Contextual Enrichment

To address the challenge of stance ambiguity in informal political discourse, we explored whether providing contextual information about users could improve classification accuracy. This approach extends the work of Malouf and Mullen (2008), who achieved 68.48% accuracy using graph-based social context (who quotes whom) combined with Naive Bayes classification. Our research investigates whether user profile summaries can provide similar contextual benefits when applied to modern LLMs. We tested seven different LLMs on the same dataset with and without user profile summaries.

5.1.1 Impact of User Profiles on Classification Accuracy

Figure 2 demonstrates that adding user profile summaries substantially enhances stance classification across all models tested. This contextual enrichment approach produced significant improvements

Model	No Context (Baseline)	With User Summaries (Enhanced)	Improvement
Grok-2-1212B	35.50%	74.00%	**+38.50%**
Meta-Llama 3.1-70B	41.50%	72.00%	**+30.50%**
Qwen 2.5-72B	37.00%	66.00%	**+29.00%**
Mistral Small-24B	34.50%	63.00%	**+28.50%**
GPT-4o Mini	34.00%	60.00%	**+26.00%**
Claude 3.7 Sonnet	42.50%	67.00%	**+24.50%**
google_gemini-2.0-flash-001	36.00%	53.50%	**+17.50%**

Figure 2: Classification accuracy comparison with and without user profile summaries.

that ranged from +17.50% to +38.50% in absolute precision.

The most striking improvement was observed with Grok-2-1212B, which saw a +38.50% increase (from 35.50% to 74.00%). Despite having a relatively low baseline performance, this model exhibited the greatest benefit from contextual information. The Meta-Llama 3.1-70B model, while starting from a higher baseline (41.50%), still achieved a substantial +30.50% improvement when provided with user summaries.

Even the model with the highest baseline accuracy, Claude 3.7 Sonnet (42.50%), gained a significant +24.50% improvement with context enhancement. Google's Gemini 2.0 Flash showed the most modest improvement at +17.50%, which aligns with a broader pattern we explore in Section 5.2.3, where we discover that models often perform sub-optimally when classifying using their own generated profiles compared to profiles generated by other models. Despite Gemini being a competent classifier overall, this particular limitation affected its performance in this experiment. To explore the maximum potential of our approach, we used all available posts except the 200 reserved for testing to generate the most comprehensive user profiles possible, which led to our peak accuracy of 74.00% with Grok-2-1212B.

Notably, our highest accuracy result (74.00% with Grok-2-1212B) surpassed the best result from Malouf and Mullen (2008) (68.48%), despite our approach using a different form of contextual information. This indicates that LLMs with user profiles can effectively leverage context in ways comparable to or better than traditional methods using explicit social network information.

5.1.2 Context Size and Selection Strategy

Our earlier experiments (Figure 3) reveal that both the quantity and selection strategy of posts used to create user profiles significantly impact classification performance. When comparing different

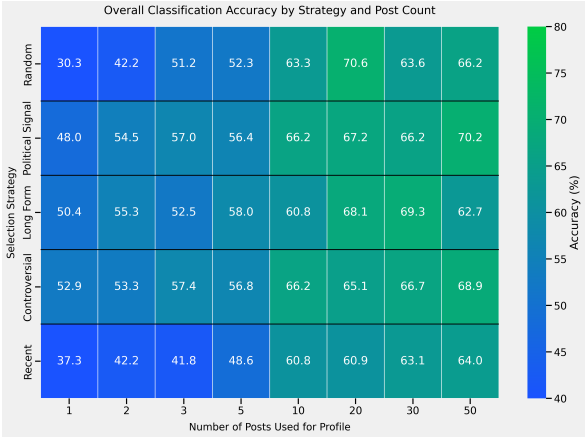

Figure 3: Accuracy by post selection strategy and number of posts used for user profiles.

post selection strategies, we found that sampling based on **political signal strength** generally outperformed other approaches, reaching 70.2% accuracy when using 50 posts per user.

However, the relationship between post count and accuracy is non-linear. We observed diminishing returns after 10-20 posts, with most strategies showing only modest gains beyond this threshold. For instance, the political signal strategy achieved 66.2% accuracy with just 10 posts, which increased only marginally to 70.2% with 50 posts.

Interestingly, the random selection strategy showed the most substantial gains when scaling from 10 posts (63.3%) to 20 posts (70.6%), suggesting that volume can partially compensate for less sophisticated selection methods. However, its performance declined with higher post counts, potentially due to the inclusion of irrelevant content that dilutes relevant signals.

These findings indicate that while providing more context generally improves performance, strategic selection of highly relevant posts yields better results than simply increasing context volume. This has important implications for real-world applications, where processing efficiency

must be balanced against classification accuracy.

5.1.3 Cross-Model Applicability

An important question is whether contextual enrichment benefits all models equally or if certain architectures are better suited to leveraging user profile information. Our experiments show that while all models improved significantly, the relative gains were inversely proportional to baseline performance. Models with weaker baseline performance (Grok, Qwen, Mistral, GPT-4o Mini) saw the largest relative improvements, suggesting that contextual information may have a normalizing effect—bringing underperforming models closer to the capabilities of stronger ones.

This pattern indicates that contextual enrichment is particularly valuable for deployment scenarios where computational constraints necessitate using smaller or less capable models. By providing well-curated user profiles, even models with limited parameters can achieve competitive stance classification performance.

5.2 Cross-Model Performance Analysis

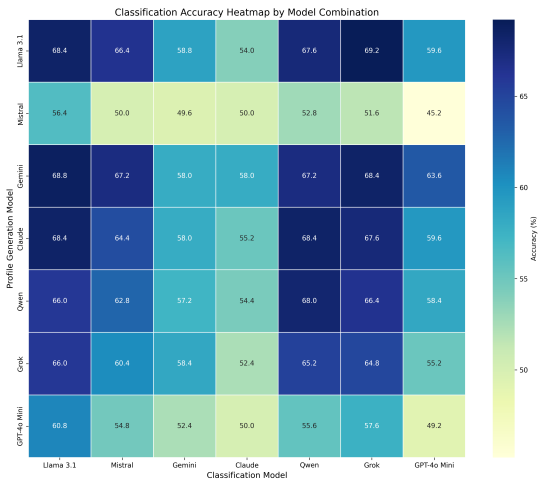

Figure 4: Classification accuracy heatmap by model combination. Profile generation models are shown on the y-axis, while classification models are on the x-axis.

To understand the relative strengths of different LLMs in the context-enriched classification pipeline, we conducted a comprehensive cross-model evaluation. As shown in Figure 4, we tested all combinations of profile generation and classification models, revealing several important patterns:

5.2.1 Profile Generation Capabilities

The vertical dimension of the heatmap reveals which models excel at generating informative user profiles. Our analysis shows that Llama 3.1, Gemini, Claude, Qwen, and Grok consistently produce high-quality profiles, enabling classification accuracies above 60% when used with strong classification models. In contrast, Mistral Small and GPT-4o Mini demonstrate weaker profile generation capabilities, with their profiles resulting in generally lower classification accuracy across all classification models. Notably, Llama 3.1 profiles yield the best overall performance, with an average accuracy of 63.4% across all classification models, suggesting superior capability in distilling relevant political patterns from user post history.

5.2.2 Classification Strengths

The horizontal dimension of the heatmap reveals which models most effectively utilize profile information for classification. Llama 3.1 and Grok stand out as the strongest classification models, achieving high accuracy regardless of which model generated the profiles. Claude and Gemini demonstrate middling performance as classifiers, while still benefiting significantly from high-quality profiles. In contrast, GPT-4o Mini consistently performs weakest as a classifier across most profile sources, suggesting potential limitations in its ability to interpret and apply contextual information.

5.2.3 Optimal Model Combinations

The most effective combinations revealed by our experiments were Gemini + Llama 3.1 (68.8% accuracy), Llama 3.1 + Grok (69.2% accuracy), and Claude + Qwen (68.4% accuracy). Interestingly, we found that most models perform better when using profiles generated by a different model rather than their own profiles (the diagonal is not consistently highest). This suggests complementary strengths between different models in the context-enriched classification pipeline. For example, while Llama 3.1 is strong in both roles, it achieves its peak performance (69.2%) when classifying posts using Grok-generated profiles rather than its own.

This finding has important practical implications, suggesting that hybrid approaches combining different models for profile generation and classification may yield better results than using a single model for the entire pipeline.

5.3 Synthesis of Findings

Our experiments reveal three key insights that advance our understanding of political stance classifi-

cation in informal discourse:

1. **Contextual enrichment significantly improves performance** across all models tested, with absolute accuracy gains of +17.50% to +38.50%. This confirms and extends Malouf and Mullen (2008)'s finding that contextual information is crucial for this task.

2. **Strategic post selection is more important than quantity** when building user profiles. The political signal selection strategy with just 10-20 posts can achieve nearly optimal performance, offering an efficient approach for real-world applications.

3. **Different models exhibit complementary strengths** in the profile generation/classification pipeline, with the best results achieved by combining models that excel in each respective role.

These findings demonstrate that modern LLMs can effectively leverage user context for political stance classification, achieving results comparable to or better than traditional methods using explicit social network information. Furthermore, our work reveals that careful optimization of contextual information and model selection can substantially enhance performance on this challenging task.

6 Conclusion

In this paper, we investigated how LLMs can be leveraged to accurately classify political stances in informal discourse by incorporating user-level contextual information. Our research demonstrates that providing summarized user profiles based on historical posts significantly enhances classification accuracy across all tested models, with improvements ranging from +17.50% to +38.50%.

We found that strategic selection of posts with strong political signals yields better results than simply maximizing context volume, with diminishing returns observed beyond 10-20 posts per user. This suggests efficient approaches for real-world applications where processing constraints may limit context size. Our cross-model evaluation further revealed that different LLMs exhibit complementary strengths in the context-enriched classification pipeline, with some models excelling at profile generation while others perform better at classification.

Our best result—74.00% accuracy with Grok-2-1212B using comprehensive user profiles—surpassed previous approaches that relied on social network information. This demonstrates that modern LLMs with appropriate contextual information can effectively address the challenge of political stance detection in informal, ambiguous discourse settings.

Limitations

While our research demonstrates significant improvements in political stance classification through contextual enrichment, several limitations should be acknowledged: (1) Our dataset from politics.com represents a specific time period and cultural context that predates current political divisions, potentially limiting direct applicability to contemporary discourse across different platforms and demographics; (2) Our LEFT/RIGHT classification framework simplifies the spectrum of political ideologies, necessary for experimental clarity but not fully reflecting the complexity of real-world political stances; (3) Practical constraints limited our testing of all possible combinations of model parameters, profile sizes, and prompt formulations. Future work could explore more nuanced political categorization beyond binary classification, test generalizability across diverse political discourse platforms, and investigate optimal context generation strategies for specific model architectures, potentially yielding even more accurate stance detection systems for real-world applications; (4) While our method focuses on user-level contextual enrichment, we did not explore reasoning-aware prompting strategies such as Chain-of-Thought (Wei et al., 2022), ReAct (Yao et al., 2023), or prefix-tuning (Li and Liang, 2021). These techniques may help models better interpret sarcastic or implicit cues in political discourse, and their integration with user-informed prompting represents a promising direction for future research.

Ethical Considerations

Our research on political stance classification raises several ethical considerations: (1) Dual-Use Potential: While intended to improve understanding of political discourse, these technologies could potentially be used for political profiling or surveillance, highlighting the importance of applications focused on enhancing communication rather than targeting individuals; (2) Algorithmic Bias: Stance classifica-

tion systems may perpetuate biases present in training data or models, necessitating monitoring for systematic errors affecting specific political groups; (3) Transparency and Consent: Applications should clearly disclose how user data is processed and political stances are inferred, with appropriate opt-out mechanisms for users whose historical data is analyzed. We recommend that implementations be accompanied by oversight mechanisms and ethical guidelines that respect political diversity and user privacy, particularly in environments where political expression may carry social or professional consequences.

References

Meta AI. 2024a. Introducing llama 3.1: Faster and stronger. Accessed: 2025-06-29.

Mistral AI. 2024b. Mistral small 24b base (2501). Accessed: 2025-06-29.

Ayah Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing user stances in political debates on twitter. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, volume 3, pages 1–20.

Emily Allaway and Kathleen McKeown. 2023. Zero-shot stance detection: Paradigms and challenges. *Frontiers in Artificial Intelligence*, 5:1070429.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2025-06-29.

Dimosthenis Antypas, Alun Preece, and Jose Camacho-Collados. 2023. Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media*, 33:100242.

Jean Aristide Aquino, Di Jie Liew, and Yung-Chun Chang. 2025. Graph-aware pre-trained language model for political sentiment analysis in filipino social media. *Engineering Applications of Artificial Intelligence*, page 110317.

Prasanta Bhattacharya, Abhijit Guha, Vidya Krishnan, Sarah Xie, and Dhanya Sridhar. 2024. Enhancing user stance detection on social media using language models: A theoretically-informed research agenda. *arXiv preprint arXiv:2502.02074*.

Josemar A. Caetano, Hélder S. Lima, Mateus F. Santos, and Humberto T. Marques-Neto. 2018. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 american presidential election. *Journal of Internet Services and Applications*, 9(1):18.

Stanley Cao and Felix Drinkall. 2024. Language models learn metadata: Political stance detection case study. *arXiv preprint arXiv:2409.13756*.

Alibaba Cloud. 2024. Qwen2.5-72b instruct. Accessed: 2025-06-29.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2024. Prompting and fine-tuning open-sourced large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Google DeepMind. 2024. Gemini 2.0 flash: Model overview. Accessed: 2025-06-29.

Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. User embedding model for personalized language prompting. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 124–131. Association for Computational Linguistics.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Margherita Gambini, Caterina Senette, Tiziano Fagni, and Maurizio Tesconi. 2024. Evaluating large language models for user stance detection on x (twitter). *Machine Learning*, 113(10):7243–7266.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and Lilian Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*, Washington, USA. AAAI Press.

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)*, pages 7360–7367, Marseille, France. European Language Resources Association.

Nayoung Kim, David Mosallanezhad, Lu Cheng, Michelle V. Mancenido, and Huan Liu. 2024. Robust stance detection: Understanding public perceptions in social media. *arXiv preprint arXiv:2309.15176*.

Takeshi Kojima, Sharan Gu, Mizuho Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4582–4597.

Robert Malouf and Tony Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research*, 18:177–190.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3):26:1–26:23.

Tony Mullen and Robert Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI spring symposium: computational approaches to analyzing weblogs*, pages 159–162.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahar Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP for Individuals (CustomNLP4U at EMNLP)*, pages 198–219.

OpenAI. 2024. Gpt-4o: Openai's new flagship model. Accessed: 2025-06-29.

Daniel Preotiuc-Pietro, Yoram Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 729–740.

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

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7370–7392.

Younes Samih and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2637–2646.

Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1413–1426. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235. Association for Computational Linguistics.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6459–6475, Miami, Florida, USA. Association for Computational Linguistics.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541, Torino, Italia. ELRA and ICCL.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Bin Wu, Zhengyan Shi, Hossein A. Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024. Understanding the role of user profile in the personalization of large language models. *arXiv preprint arXiv:2406.17803*.

xAI. 2024. Introducing grok-2. Accessed: 2025-06-29.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Chenchen Ye, Linhai Zhang, Yulan He, Deyu Zhou, and Jie Wu. 2021. Beyond text: Incorporating metadata and label structure for multi-label document classification using heterogeneous graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3162–3171, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiarui Zhang. 2024. Guided profile generation improves personalization with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do llms recognize your preferences? evaluating personalized preference following in llms. In *International Conference on Learning Representations (ICLR)*. Oral Presentation.

Zhiwei Zhou and Erick Elejalde. 2024. Unveiling the silent majority: stance detection and characterization of passive users on social media using collaborative filtering and graph convolutional networks. *EPJ Data Science*, 13(28).

A User Context and Profile Summarization

A.1 User Profile Summarization Prompt

Analyze the following set of forum posts by the user and create a concise political profile summary. For this task:

1. Identify any consistent political indicators in their posts (criticism of specific politicians/parties, stance on issues, etc.)
2. Note recurring topics this user discusses
3. Observe distinctive language patterns (formal/informal, emotional/detached, specific phrases)
4. Identify who/what they consistently criticize or support
5. Determine if there's sufficient evidence to classify them as LEFT, RIGHT, or UNKNOWN

Format your response as a JSON object with these fields:

```
{
  "username": "the username",
  "political_leaning": "left/right/unknown",
  "confidence": "high/medium/low",
  "key_indicators": ["3-5 specific examples from posts that indicate political leaning"],
  "recurring_topics": ["list frequent topics"],
  "language_style": "brief description of their communication style",
  "sentiment_patterns": "who/what they criticize or support",
  "context_notes": "any additional relevant information"
}
```

IMPORTANT:

- Focus on clear patterns rather than isolated statements
- Maintain objectivity and avoid over-interpreting ambiguous content
- If there isn't sufficient evidence to determine orientation, mark as "unknown"
- Ensure your response is a valid JSON object

A.2 Classification with Profile Summary Prompt

Analyze the following discussion group post and classify the author's political orientation.

IMPORTANT CONTEXT ABOUT THIS USER:
{profile_summary}

Take the above user profile into account when analyzing this post. The profile reflects patterns from the user's previous posts, which may provide context for this specific post.

Provide your response in this exact JSON format:

```
{
  "orientation": "LEFT|RIGHT|UNKNOWN",
  "explanation": "A detailed explanation of why you chose this classification based on the content"
}
```

B Post Selection Strategy Implementation Details

In this section, we provide the detailed implementation of our post selection strategies, particularly the **PoliticalSignalSelection** algorithm that performed best in our experiments.

B.1 PoliticalSignalSelection Algorithm

The **PoliticalSignalSelection** strategy uses a weighted lexicon approach to identify posts with strong political content. The algorithm works as follows:

1. **Term Weighting:** Political terms are categorized and weighted based on their signal strength:
   - *General political terms* (weight 1): 'politics', 'political', 'government', 'policy', 'policies', 'election', 'vote', 'voting', 'democracy', 'democratic'
   - *Party-specific terms* (weight 2): 'democrat', 'democratic party', 'liberal', 'progressive', 'socialism', 'left', 'left-wing', 'republican', 'gop', 'conservative', 'right', 'right-wing', 'trump', 'biden', 'obama', 'maga', 'tea party'

- *Hot-button issues* (weight 3): 'abortion', 'gun', 'immigration', 'climate', 'tax', 'healthcare', 'obamacare', 'socialism', 'vaccine', 'blm', 'black lives matter', 'defund', 'wall', 'border'

2. **Post Scoring:** For each post:
   - Count occurrences of each political term in the post text
   - Multiply each term's count by its assigned weight
   - Sum these weighted counts to calculate the post's political signal score
   - Add a small random factor (0-0.01) to break ties between posts with identical scores
   - Apply a +5 point boost to posts from explicitly political subforums
3. **Post Selection:** After scoring all posts:
   - Sort posts by their political signal scores in descending order
   - Select 60% of the required posts from those with highest scores
   - Select the remaining 40% to ensure topic diversity, prioritizing posts with different term distributions

This algorithm effectively identifies posts with strong political indicators while maintaining sufficient topical diversity in the selected content for user profile generation.

C Additional Figures

This appendix contains larger versions of the figures presented in the main text, allowing for more detailed examination.

Model	No Context (Baseline)	With User Summaries (Enhanced)	Improvement
Grok-2-1212B	35.50%	74.00%	**+38.50%**
Meta-Llama 3.1-70B	41.50%	72.00%	**+30.50%**
Qwen 2.5-72B	37.00%	66.00%	**+29.00%**
Mistral Small-24B	34.50%	63.00%	**+28.50%**
GPT-4o Mini	34.00%	60.00%	**+26.00%**
Claude 3.7 Sonnet	42.50%	67.00%	**+24.50%**
google_gemini-2.0-flash-001	36.00%	53.50%	**+17.50%**

Figure 5: Larger version of Figure 2: Classification accuracy comparison with and without user profile summaries.

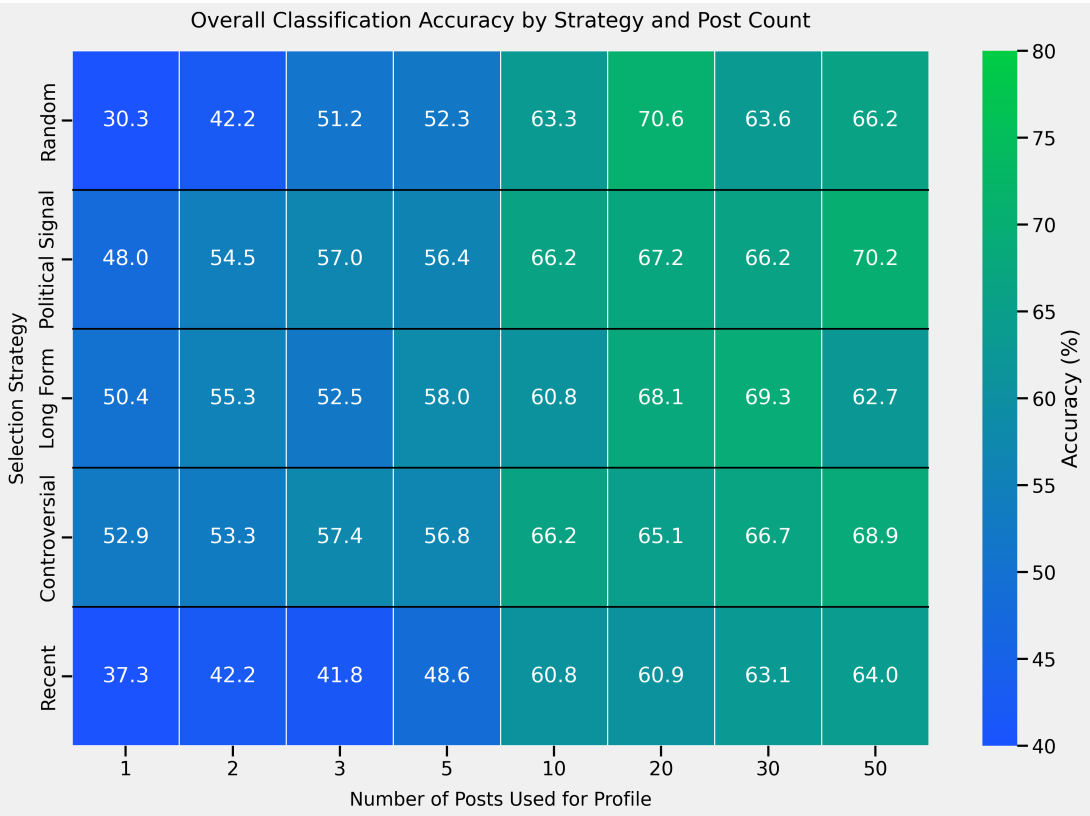

Figure 6: Larger version of Figure 3: Accuracy by post selection strategy and number of posts used for user profiles.

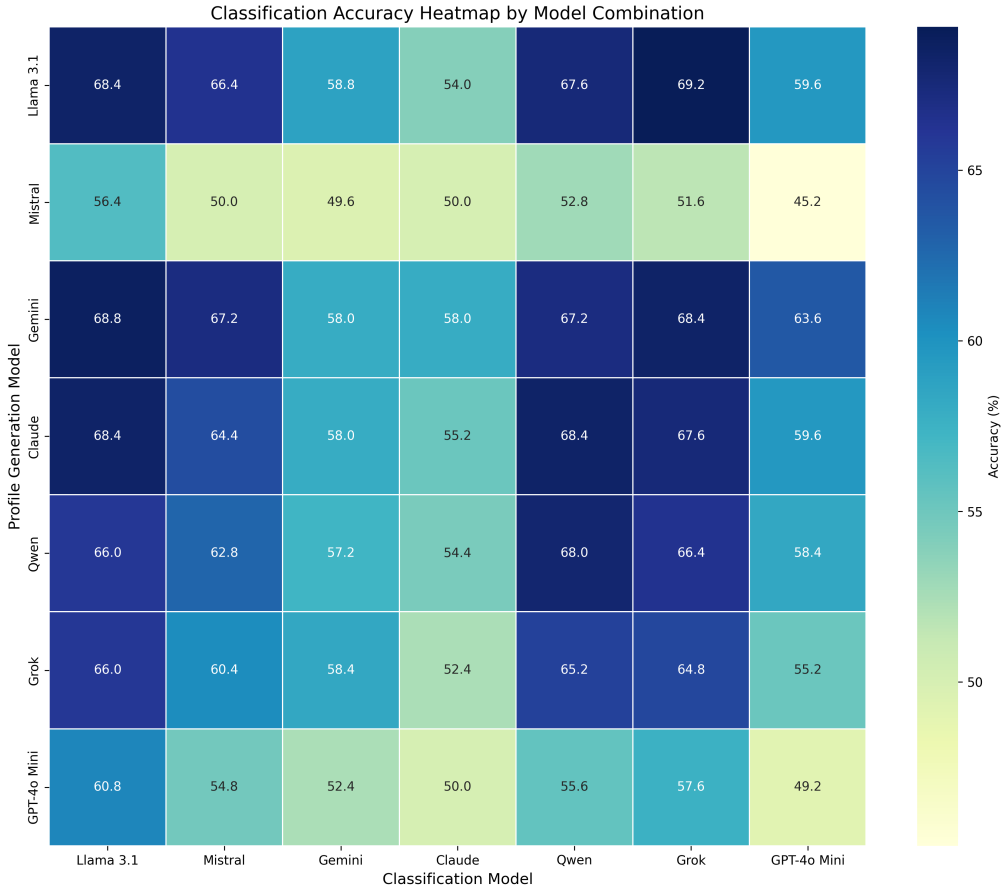

Figure 7: Larger version of Figure 4: Classification accuracy heatmap by model combination.