Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Political Stance Detection on X During the 2024U.S. Election: A Comparative Study of Classical, Neural, and Transformer Models

Qile Wang

kylewang@udel.edu

University of Delaware

Sahar Ostadrahimi

University of Delaware

Safoura Faghri

University of Delaware

Mohammad Baksh

University of Delaware

Matthew Louis Mauriello

University of Delaware

Kenneth E. Barner

University of Delaware

Research Article

Keywords: Social Media, Machine Learning, Benchmark, Computational Social Science

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9735442/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Political Stance Detection on X During the 2024 U.S. Election: A Comparative Study of Classical, Neural, and Transformer Models

Qile Wang[1*†], Sahar Ostadrahimi[1†], Safoura Faghri[1†], Mohammad Baksh[1], Matthew Louis Mauriello[1], Kenneth E. Barner[1]

[1*]University of Delaware, Newark, Delaware, United States.

*Corresponding author(s). E-mail(s): kylewang@udel.edu;
Contributing authors: sorahimi@udel.edu; safouraf@udel.edu; mmbaksh@udel.edu; mlm@udel.edu; barner@udel.edu;
†These authors contributed equally to this work.

Abstract

As online social networks (OSNs) play an increasingly influential role in political communication, accurately detecting political stance from noisy, short-form content remains a significant challenge in computational social science (CSS). This paper presents a comparative multi-feature framework for political stance detection during the 2024 U.S. presidential election, focusing on candidate affiliation toward Donald Trump and Joe Biden–Kamala Harris. Using nearly 100,000 posts from X, we introduce a candidate affiliation feature that extends beyond coarse sentiment polarity and leverages topic modeling to support scalable annotation and thematic characterization. Under a unified preprocessing, labeling, and evaluation pipeline, we compare classical machine learning models, sentence-embedding-based neural architectures, and a domain-adapted transformer. Results show that appropriate preprocessing and task-aligned text representation often exert greater influence on performance than model complexity alone. While fine-tuned BERTweet achieves the strongest in-domain performance, carefully engineered classical models, particularly linear support vector machines, remain highly competitive while offering substantial advantages in computational efficiency, interpretability, and reproducibility. Multi-feature and ablation analyses further indicate that expanded metadata provides only modest and inconsistent gains beyond strong text representations. Overall, this

work offers practical guidance for designing scalable, interpretable, and resource-efficient political stance detection systems while emphasizing the importance of cautious, deployment-aware evaluation under dataset-specific conditions.

Keywords: Social Media, Machine Learning, Benchmark, Computational Social Science

1 Introduction

Online social networks (OSNs) have become central infrastructures for political communication, public discourse, and opinion formation in modern democratic societies (Brena et al. 2019). Among these platforms, X (formerly Twitter) plays a particularly influential role due to its real-time information exchange, public accessibility, and global reach. During national elections, X serves not only as a mechanism for disseminating information but also as a large-scale behavioral environment in which political preferences, ideological alignments, and collective narratives are continuously expressed, amplified, and contested.

The scale, immediacy, and public availability of X data have made the platform an important resource for computational social science (CSS), political communication research, and natural language processing. Prior studies demonstrate that social media activity and sentiment can correlate with voter behavior, candidate support, and broader electoral dynamics, suggesting that online discourse may serve as a proxy for public political attitudes under certain conditions (Chaudhry et al. 2021; Kumari and Singh 2024; Alvi et al. 2023). Analyses of election cycles further indicate that temporal shifts in social media discourse often align with changes in political confidence, polarization, and campaign momentum (Chaudhry et al. 2021; Kumari and Singh 2024). Similar findings across regional and international contexts reinforce the value of X as a scalable source for political discourse analysis (Alvi et al. 2023; Kowsik et al. 2024).

Despite this promise, robust political stance detection on X remains methodologically challenging. Posts are typically short, informal, and weakly structured, often containing slang, abbreviations, emojis, hashtags, sarcasm, and rapidly evolving political language. These characteristics complicate semantic interpretation and can obscure ideological intent. In addition, misinformation, irony, and implicit rhetorical framing frequently decouple expressed stance from surface sentiment, reducing the reliability of sentiment-only approaches. Recent surveys consistently emphasize that these linguistic and contextual complexities constrain both predictive accuracy and interpretability in social media analysis (Alvi et al. 2023; Kowsik et al. 2024; Albladi et al. 2025; Skoularikis et al. 2026).

Current research also exhibits several practical limitations, as suggested by an interview-based case study on current OSN data analysis (Wang et al. 2025). First, much of the literature relies heavily on coarse sentiment polarity, which often fails to capture politically meaningful distinctions between affective tone and candidate alignment. Political stance is frequently conveyed through indirect signaling, rhetorical

framing, or issue positioning rather than explicit positive or negative sentiment, making sentiment classification an incomplete proxy for political orientation (Alvi et al. 2023; Kowsik et al. 2024; Albladi et al. 2025). Second, while classical machine learning (ML) approaches remain computationally efficient, they can struggle with noisy, heterogeneous, and highly contextualized political discourse (Aamir et al. 2024; Qi and Shabrina 2023; Rustam et al. 2021). Conversely, deep learning and transformer-based architectures often improve contextual understanding but require substantial computational resources, larger labeled datasets, and reduced interpretability, limiting their practicality for transparent or real-time deployment (Albladi et al. 2025; Marshan et al. 2025). Third, existing studies rarely provide unified comparative evaluations across classical, neural, and transformer paradigms under controlled preprocessing and feature conditions, making it difficult to assess trade-offs among predictive performance, computational efficiency, and deployment feasibility.

To address these gaps, this work develops and evaluates a comparative multi-feature framework for political stance detection on X during the 2024 U.S. presidential election. Building on stance detection principles (Küçük and Can 2022), we introduce a candidate-affiliation feature that captures user alignment with major presidential candidates beyond coarse sentiment polarity. This representation is paired with thematic characterization through topic modeling and evaluated within a unified preprocessing, labeling, and experimental framework. We systematically compare classical machine learning models, sentence-embedding-based neural architectures, and a domain-adapted transformer to examine how preprocessing rigor, feature representation, and model complexity jointly affect predictive performance, interpretability, and computational cost. The primary contributions of this work are as follows:

- We introduce a candidate affiliation feature that models political alignment toward Donald Trump versus Joe Biden–Kamala Harris, providing a stance-oriented representation beyond conventional sentiment polarity.
- We integrate topic modeling techniques, including LDA-based thematic analysis, to support candidate affiliation annotation and enhance the interpretability of the structure of political discourse.
- We conduct a systematic comparative evaluation of classical machine learning, neural, and transformer-based models under consistent preprocessing and evaluation protocols, demonstrating that carefully engineered classical approaches can remain competitive despite substantially lower computational complexity.
- We provide practical design insights for scalable political stance detection by analyzing trade-offs among predictive accuracy, interpretability, feature complexity, and deployment efficiency relevant to election monitoring systems.

2 Related Work

This section reviews prior research on computational methods for political discourse analysis in online social networks (OSNs) and situates stance detection as a related but distinct analytical task. We first examine the evolution of machine learning approaches

Table 1: Representative review of machine learning approaches for OSN discourse analysis, emphasizing X/Twitter-based political and social studies. Reported F1-score or accuracy (%) values reflect the best-performing model in each study and are provided for contextual reference, as datasets and evaluation protocols vary. Abbreviations: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB/XGBoost), Feedforward Neural Network (FNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU).

Authors	Platform	Size	Task	Methods	F1/Acc.
Chaudhry et al. (2021)	X	18M	Election sentiment	NB, LR, SVM	0.95
Rustam et al. (2021)	X	7.5K	COVID sentiment	RF, XGB, SVM, LSTM	0.93
Pandya et al. (2021)	X	50K	Consumer sentiment	SVM, NB, BiLSTM	0.88
Qi and Shabrina (2023)	X	15K	Political sentiment	NB, SVM, Lexicon	0.71
Mansur et al. (2023)	X	10K–100K	Hate speech (survey)	SVM, LR, RF, CNN, LSTM, Transf.	0.92
Alturayeif et al. (2023)	X	5K–50K	Stance detection (survey)	SVM, LR, CNN, LSTM, Attn., Transf.	0.90
Alvi et al. (2023)	X	10K–2M	Election prediction (survey)	NB, SVM, RF, LSTM, BERT	95%
Aamir et al. (2024)	X	50K	Sentiment classification	SVM, NB, LSTM, CNN	0.93
Kowsik et al. (2024)	X	100K	Bias detection	SVM, Topic Modeling	0.83
A. Semary et al. (2024)	X, Amazon	25K	Feature representation	RF	99%
Papadimitriou et al. (2024)	X, Reddit	100K	Cross-platform sentiment	SVM, LR, CNN	0.90
Kumari and Singh (2024)	X	3M	Election prediction	RF, NB	92%
Marshan et al. (2025)	X, Reddit	60K	Hate speech severity	LR, SVM, CNN, LSTM	0.92
Albladi et al. (2025)	X	200K	Polarization detection	LSTM, GRU, CNN	0.93
Jonnala et al. (2025)	X	60K	Political sentiment	BiLSTM, LR	0.83
Mohammed Alsekait et al. (2025)	Facebook	30K	Social sentiment	LR, SVM, RF, FNN	99%
Padalko et al. (2025)	X	42K	Pro-Russian Classification	LR, SVM, BiLSTM, KNN, XGB	82%

applied to political content on X, emphasizing shifts in feature representation, methodological complexity, and deployment considerations. We then review stance detection literature more directly and conclude by identifying key methodological gaps that motivate the proposed comparative framework.

2.1 Computational Methods for Political Discourse in OSNs

Social media platforms, and X in particular, have long been recognized as large-scale observational environments for studying political behavior, voter alignment, and public opinion. Early work demonstrated that relatively simple behavioral indicators, including party mentions, hashtags, and retweet activity, could be highly predictive of political affiliation. For example, Boutet et al. (2012) showed that party-reference frequencies during the 2010 UK General Election achieved approximately 86% accuracy without supervised learning, illustrating that lightweight, interpretable signals can reveal substantial political structure. Beyond electoral affiliation, subsequent studies have also used X to examine sentiment, engagement, and discourse surrounding broader social issues, demonstrating that online interaction patterns can expose nuanced public attitudes across domains (Mullins et al. 2025).

To contextualize methodological trends, Table 1 provides a targeted synthesis of representative machine learning studies relevant to political discourse analysis on OSN data. Rather than serving as an exhaustive survey, this review emphasizes influential studies spanning the past decade, selected for their methodological relevance, supervised-learning focus, and reported evaluation performance. Across this literature, classical machine learning approaches—including Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Random Forests—remain widely adopted and often achieve strong predictive performance on noisy, medium-scale datasets. Reported F1 scores frequently exceed 90% under well-structured task conditions. Neural architectures, including CNNs, LSTMs, GRUs, and hybrid models, have increasingly been applied to more complex tasks such as detecting political polarization, assessing hate speech severity, and classifying cross-platform sentiment. However, performance gains over classical baselines are often inconsistent and highly dependent on dataset size, task design, and preprocessing rigor. Considerable variation in platform, domain, and feature construction further limits direct comparability across studies, underscoring the fragmented nature of the broader literature.

A substantial portion of prior work has focused specifically on sentiment analysis as a proxy for political attitudes and election outcomes. Studies examining U.S. presidential election cycles have reported correlations between aggregated X sentiment and voter behavior, reinforcing sentiment polarity as a foundational feature in election-related OSN research (Chaudhry et al. 2021; Rustam et al. 2021). While these studies provide important evidence that social media sentiment can reflect political dynamics, they typically rely on coarse sentiment categories and treat posts primarily as isolated textual units. As a result, they often fail to capture broader behavioral, temporal, or contextual dimensions of political communication. To improve robustness under noisy conditions, many researchers adopted classical supervised approaches paired with bag-of-words or TF–IDF representations. Comparative evaluations consistently identify SVM-based models as particularly effective for short, noisy political text and frequently competitive with more complex architectures on moderate-scale datasets (Qi and Shabrina 2023; Aamir et al. 2024). Nonetheless, these models often overlook non-textual signals, such as engagement dynamics, misinformation cues, and content moderation indicators, thereby constraining interpretability and limiting explanatory depth.

Deep learning approaches were later introduced to better capture nonlinear semantic patterns and contextual dependencies. Feedforward neural networks, CNNs, LSTMs, and GRUs have been widely applied to political sentiment, ideological classification, and hate speech detection, often outperforming classical approaches when sufficiently large labeled datasets are available (Patel et al. 2024; Albladi et al. 2025; Marshan et al. 2025). At the same time, multiple studies report diminishing returns from architectural complexity alone, particularly when preprocessing and feature engineering are carefully optimized (Mohammed Alsekait et al. 2025; Pandya et al. 2021). More recently, transformer-based architectures such as BERT, BERTweet, and hybrid BERT-based models have established state-of-the-art performance on X by leveraging contextualized language representations and large-scale pretraining. Fine-tuned transformers frequently exceed 90% performance on political sentiment and stance-related tasks, particularly in domain-adapted settings (Papadimitriou et al. 2024; Albladi et al. 2025). However, these improvements often come at the expense of computational efficiency, transparency, and suitability for real-time or resource-constrained deployment.

Parallel research directions have emphasized behavioral, network, and temporal dimensions of political discourse. Retweet graphs, mention networks, and interaction structures reveal strong ideological homophily and polarization, enabling inference of political affiliation from relational patterns independent of textual sentiment (Conover et al. 2011; Boutet et al. 2012). Temporal analyses further demonstrate that political discourse evolves dynamically across campaign phases, debates, and major political events, suggesting that static text representations alone may fail to capture meaningful longitudinal variation. Despite these advances, recent reviews consistently identify fragmentation as a central limitation in the field (Alvi et al. 2023; Kowsik et al. 2024). Many studies focus narrowly on isolated tasks such as sentiment analysis, misinformation detection, or hate speech classification without integrating textual, behavioral, temporal, and representational dimensions into a unified comparative framework. Even when richer feature sets are introduced, their contributions are rarely evaluated systematically across modeling families.

Building on these gaps, the present work develops a comparative multi-feature framework for political stance detection during the 2024 U.S. presidential election. By combining candidate-affiliation labeling, thematic topic structure, multiple model families, and controlled feature ablation within a unified preprocessing and evaluation pipeline, this study seeks to provide a more rigorous assessment of how feature design, model complexity, and computational trade-offs jointly influence scalable, interpretable political discourse analysis on X.

2.2 OSN Stance Detection

Stance detection has emerged as a central task in computational social science (CSS) and natural language processing, driven by the rapid expansion of user-generated content on online social networks (OSNs). Unlike sentiment analysis, which primarily measures affective polarity, stance detection seeks to infer an author's position toward a specific target, such as a political figure, policy issue, or sociopolitical event (Küçük and Can 2020). Common stance categories include *Favor*, *Against*, and *Neither*,

although task-specific taxonomies frequently vary depending on domain and application (Alturayeif et al. 2023). This distinction is particularly important in political discourse, where sentiment and stance may diverge substantially; users may express negative sentiment while still signaling political support, or vice versa.

Early stance detection research primarily relied on traditional supervised machine learning approaches, including Support Vector Machines (SVMs) and Logistic Regression combined with lexical representations such as n-grams and TF–IDF (Küçük and Can 2020; Darwish et al. 2020). These studies demonstrated that stance-related signals often emerge through recurring lexical patterns, ideological framing, and topic-specific language. However, survey literature consistently identifies persistent challenges in modeling implicit stance, sarcasm, rhetorical indirection, and domain shift across topics, communities, and time periods (Alturayeif et al. 2023). Such challenges are particularly pronounced in political OSN environments, where users frequently encode ideological alignment indirectly through slogans, framing, or engagement behavior rather than explicit declarative statements.

To address these limitations, more recent work has increasingly adopted deep learning and transformer-based architectures. Recurrent neural networks, attention-based models, and fine-tuned transformer systems improve contextual understanding by leveraging distributed semantic representations and large-scale pretraining (Chauhan et al. 2019; Zhang 2025; Haouari and Elsayed 2024). These models have demonstrated improved performance on many stance-related tasks, particularly when sufficient labeled data and domain adaptation are available. However, their gains often come with increased computational demands, reduced interpretability, and greater dependence on large annotated corpora, raising important concerns regarding scalability, transparency, and real-time applicability (Alturayeif et al. 2023).

Beyond text-only approaches, several studies emphasize that stance is frequently expressed through broader behavioral and relational structures. User interaction patterns, engagement behaviors, and social context can reveal stance even when it is not explicitly stated in linguistic content (Aldayel and Magdy 2019). Discourse-aware frameworks further conceptualize stance as relational, where posts position themselves relative to prior claims through support, denial, or questioning (Zubiaga et al. 2018). These perspectives suggest that stance is often embedded within larger conversational, behavioral, and temporal ecosystems rather than in isolated textual units. Accordingly, effective stance detection may require integrating textual semantics with broader contextual signals.

Despite substantial progress, several gaps remain across both stance detection and political discourse analysis. First, much of the literature evaluates isolated modeling paradigms or narrow feature sets, limiting broader understanding of trade-offs among predictive accuracy, computational efficiency, interpretability, and deployment feasibility. Second, the relative contributions of textual, behavioral, temporal, and engagement-related features are rarely systematically examined within a unified comparative framework. Third, many studies emphasize headline predictive performance without adequately evaluating robustness, feature dependence, or practical constraints relevant to scalable election-monitoring systems.

This work addresses these gaps by developing a comparative multi-feature framework for political stance detection on X during the 2024 U.S. presidential election. We introduce a candidate-affiliation feature that extends beyond coarse sentiment categories to better capture politically relevant alignment and systematically evaluate its interactions with linguistic, behavioral, and temporal signals. By comparing classical machine learning models, sentence-embedding-based neural architectures, and transformer-based approaches under identical preprocessing and evaluation conditions, this study provides a feature-centric assessment of performance trade-offs across modeling families. This positioning directly motivates the methodological framework described in the following section.

3 Methodology

This section presents the proposed comparative framework for political stance detection on X during the 2024 U.S. presidential election. The methodological objective is to enable a controlled evaluation across multiple modeling paradigms while systematically isolating the effects of preprocessing, feature representation, and model complexity on predictive performance, robustness, and generalization.

We first describe the dataset and preprocessing pipeline used to transform noisy, weakly structured social media posts into a standardized analytical corpus. We then detail the feature engineering framework, including the construction of linguistic, behavioral, temporal, engagement, and candidate affiliation features designed to capture complementary dimensions of political discourse. Finally, we present the modeling approaches evaluated in this study, spanning classical machine learning baselines, sentence-embedding-based neural architectures, and a domain-adapted transformer. Fig. 1 summarizes the complete preprocessing, feature engineering, labeling, and modeling workflow. To ensure fair, reproducible, and interpretable comparison, all models are evaluated under identical data splits, preprocessing conditions, and performance protocols.

3.1 Datasets

We utilize the **USE24-XD** dataset, which provides full content from Wang et al. (2026), originally developed for large-scale analysis of harmful social media content during the 2024 U.S. election cycle. The dataset contains nearly 100,000 posts collected between October 17, 2024, and July 16, 2025, spanning pre-election, election, and post-election phases of the 2024 U.S. presidential cycle. This temporal coverage enables analysis across multiple stages of political discourse, capturing both campaign-period dynamics and post-election reactions.

3.1.1 Data Prepossessing

Political discourse on X is inherently noisy, informal, and weakly structured, creating substantial challenges for reliable computational analysis. Posts frequently include slang, abbreviations, emojis, hashtags, sarcasm, misspellings, and other platform-specific conventions that can obscure political stance, semantic intent, and behavioral

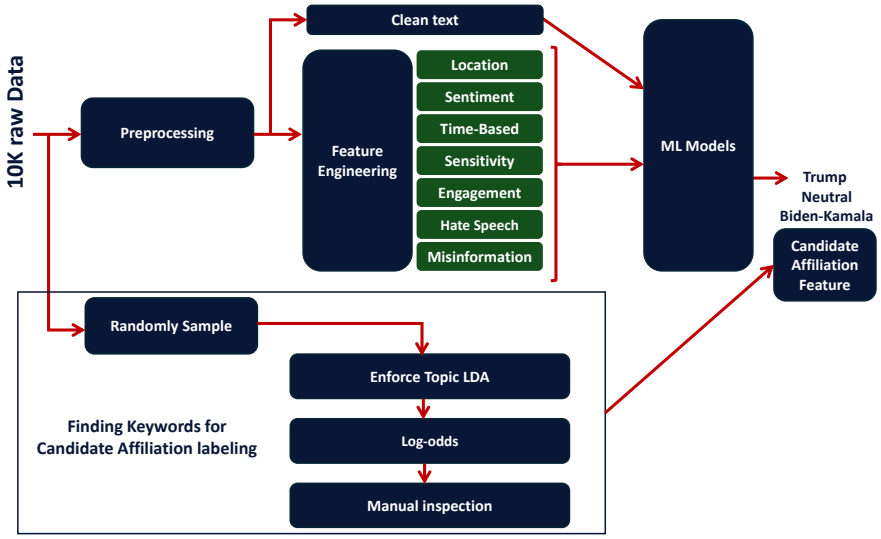


Fig. 1: End-to-end preprocessing, feature engineering, candidate affiliation labeling, and modeling pipeline for political stance detection on X.

signals. To improve analytical reliability, reproducibility, and model stability, we designed a rigorous multi-stage preprocessing pipeline that transforms raw posts into a standardized yet information-preserving representation, as illustrated in Fig. 2.

Because all downstream models, linguistic resources, and sentiment tools used in this study operate in English, we first restricted the corpus to English-language posts using platform-provided language metadata. This filtering step reduces linguistic heterogeneity, ensures compatibility with tokenization and lemmatization resources, and establishes a consistent foundation for downstream feature extraction and comparative evaluation.

We then applied a sequence of targeted normalization procedures to reduce noise while preserving politically salient information. Repost markers were removed to eliminate duplicate semantic content, and URLs were stripped using regular expression filtering. Mentions were processed selectively: references to political figures, campaign entities, or election-relevant actors were replaced with standardized tokens, while non-political mentions were removed to reduce lexical sparsity. Hashtags were converted into lexical text through underscore splitting and camelCase decomposition, preserving campaign slogans, issue references, and partisan phrases as semantically meaningful tokens. Emojis and emoticons were mapped to sentiment-relevant textual tokens to preserve affective information, consistent with prior evidence that emoji usage can encode meaningful linguistic, emotional, and sociocultural signals (Kejriwal et al. 2021). All text was lowercased, punctuation was removed, and whitespace was normalized.

To further reduce lexical variability, we applied a custom political lexicon to correct frequent misspellings and normalize politically salient terminology. Tokens were

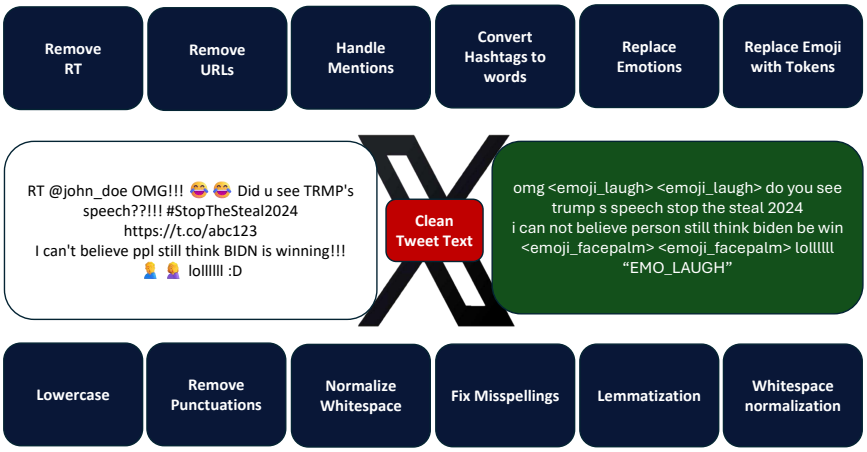


Fig. 2: Text preprocessing pipeline applied to raw X posts prior to feature extraction, candidate affiliation labeling, and downstream modeling.

then lemmatized to their base forms, retaining alphabetic terms and custom political markers while preserving pronouns to maintain key grammatical structure. Posts that became empty after preprocessing were discarded. To reduce low-information noise, posts with fewer than three remaining tokens were removed unless they included at least one predefined political keyword, thereby preserving short yet potentially meaningful content, such as slogans, endorsements, or candidate references. Finally, duplicate posts were removed when both the author identifier and the cleaned text were identical, ensuring that the final dataset more accurately reflected distinct communicative instances.

After preprocessing and filtering, the dataset was reduced from approximately 100,000 raw posts to roughly 90,000 high-quality posts. This pipeline was designed to balance aggressive noise reduction with preservation of subtle stance-relevant cues and was applied consistently across all evaluated modeling frameworks to ensure fair comparative analysis.

3.1.2 Feature Engineering

Following preprocessing, as illustrated in Fig. 1, each post was represented using a structured feature set designed to capture complementary dimensions of political discourse, including linguistic expression, affective tone, behavioral engagement, temporal context, and platform metadata. Feature construction was informed by prior work in computational social science (CSS) and political communication, with emphasis on interpretability, robustness, and reproducibility under noisy social media conditions.

To characterize affective content, each post was assigned both a continuous sentiment score in the range $[-1, 1]$ and a categorical sentiment label (*positive*, *negative*,

or *neutral*), enabling representation of both graded emotional tone and discrete polarity. To capture communicative dimensions beyond sentiment, hate-related keyword frequencies were used to derive a binary hate flag, while keyword-based indicators of misinformation were used to construct a binary misinformation flag.

Temporal context was encoded using hour-of-day, day-of-week, and election-phase indicators spanning pre-election, election-day, and post-election periods. Behavioral engagement was represented using aggregate interaction metrics, including likes, reposts, replies, quotes, and bookmarks, normalized by impressions to compute engagement rate and log-transformed to reduce heavy-tailed variance. A platform-provided binary sensitivity flag was also retained to capture moderation-relevant content.

Geographic context was derived from user-provided free-text location fields, which were cleaned and normalized to indicate whether a post originated from the United States, whether a valid U.S. state could be inferred, and whether the location field appeared missing or unreliable. When available, state identity was encoded using one-hot representations.

Together, these features provide a compact multidimensional representation of political communication that extends beyond text alone, forming the basis for candidate affiliation annotation, downstream modeling, and feature ablation analyses.

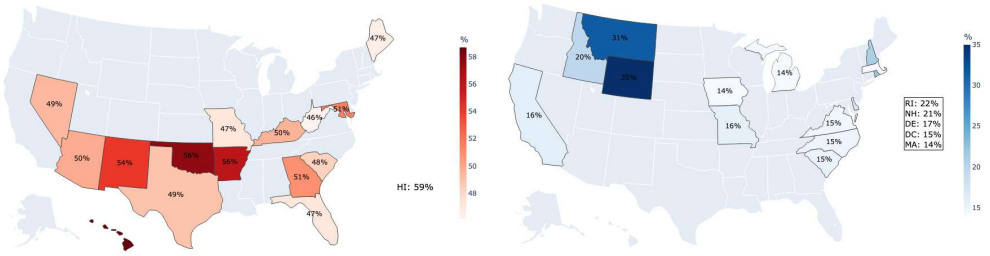
3.1.3 Candidate Affiliation Annotation

Candidate affiliation was modeled as a three-class variable: *Trump*, *Biden–Kamala*, and *Neutral*. Labels were assigned using a rule-based heuristic pipeline, summarized in Fig. 1, and refined through multiple validation stages. Initial keyword rules based on candidate names, aliases, and campaign-specific terminology were first applied to a seed corpus. Unsupervised topic modeling, including standard LDA and guided seed-word LDA, was then used to identify salient lexical patterns, refine candidate-relevant themes, and remove ambiguous or noisy terms. Log-odds ratios were subsequently computed to identify keywords that were statistically discriminative across partisan groups.

To validate and improve the labeling process, a random sample of 200 posts was manually annotated by two researchers to establish a reference ground truth. Annotation discrepancies were reviewed and resolved collaboratively, and keyword lists were iteratively refined by removing weak indicators and incorporating nuanced stance expressions identified during manual inspection. Political communication insights and GPT-assisted qualitative review were used as auxiliary tools to improve coverage and interpretability.

We evaluated agreement between manual annotation and the refined keyword-based labeling framework using inter-rater reliability (IRR) metrics. The resulting Cohen's Kappa of 0.90 and percentage agreement of 94% indicate strong to near-perfect agreement according to McHugh (2012). The finalized annotation procedure was then applied to the full dataset. In the resulting corpus, 48% ($n = 27{,}340$) of posts were labeled as *Trump*, 40% as *Biden–Kamala*, and 12% as *Neutral*. Figure 3 further illustrates geographic variation in candidate affiliation using available location proxies.

Because political alignment on social media is often implicit, sarcastic, and context-dependent, multiple refinement cycles were necessary to balance label precision with



(a) Top 15 states by *Trump* support. (b) Top 15 states by *Biden–Kamala* support.

Fig. 3: Geographic distribution of candidate-aligned support across U.S. states based on available location proxies. The left panel highlights the 15 states with the highest percentage of *Trump* affiliation posts, while the right panel highlights the 15 states with the highest percentage of *Biden–Kamala* affiliation posts. Percentages are computed within each state across posts labeled as *Trump*, *Biden–Kamala*, or *Neutral*.

coverage. Although heuristic, this annotation framework provides a scalable, interpretable, and practically reproducible method for large-scale candidate affiliation labeling in politically oriented OSN discourse.

3.2 Stance Detection

We formulate political stance detection as a three-class classification problem, distinguishing among *Trump*, *Biden–Kamala*, and *Neutral* posts. To enable a rigorous comparative evaluation across modeling paradigms, we adopt a staged framework spanning interpretable classical machine learning baselines, neural architectures built on learned sentence representations, and a domain-adapted transformer model. All approaches are trained and evaluated under identical preprocessing pipelines, data splits, and evaluation protocols to ensure that observed performance differences primarily reflect representational and architectural choices rather than experimental inconsistencies. The dataset is split into 64% training, 16% validation, and 20% test sets with strict stratification. The following subsections describe each model family in increasing order of complexity, as summarized in Table 2. Exact hyperparameter values for all optimal models are provided in the released code repository.

3.2.1 Traditional ML

Classical machine learning models provide computationally efficient and interpretable baselines for medium-scale political text classification. Cleaned posts are represented using unigram and bigram TF–IDF features, producing sparse vectors of approximately 20,000–30,000 dimensions that capture politically salient lexical patterns, including candidate references, slogans, and recurring partisan framing.

We first evaluate multinomial Logistic Regression as a probabilistic linear baseline, optimized using L2-regularized cross-entropy loss with the L-BFGS solver. The regularization strength is selected based on the validation macro-F1. We also evaluate a linear Support Vector Machine (SVM) with a one-versus-rest formulation and squared

Table 2: Optimal model hyperparameter configurations based on validation macro-F1.

Model	Hyperparameter	Value / Configuration
Logistic Regression	Max features	20,000
	Penalty	L2 regularization
	Solver	L-BFGS
	Regularization (C)	$C = 0.3$
Linear Support Vector Machines (SVM)	Max features	20,000
	Kernel	Linear
	Loss	Squared hinge loss
	Multiclass strategy	One-vs-rest
	Regularization (C)	$C = 1.0$
Histogram Gradient Boosting (HGB)	Max features	30,000, Truncated SVD features 100
	Loss	Multiclass log loss
	Learning rate	0.1
	Iterations	200
Sentence-BERT (SBERT) + Multilayer Perceptron (MLP)	Semantic embeddings	Frozen all-mpnet-base-v2 embeddings (768-dim)
	Architecture	Two fully connected hidden layers: 256, 128
	Activation & normalization	GELU activation with batch normalization
	Regularization	Dropout = 0.5
	Optimizer & loss	AdamW optimizer, Multiclass cross-entropy loss
	Learning rate	5×10^{-4}
	Weight decay	1×10^{-3}
	Batch size	128
BERTweet	Max sequence length	128
	Batch size	16
	Optimizer	AdamW
	Loss	Multiclass cross-entropy
	Learning rate	2×10^{-5}
	Weight decay	0.01
	Training epochs	4
	Scheduler	Linear warmup/decay with 10% warmup
	Gradient clipping	1.0

hinge loss. Due to its robustness in high-dimensional sparse feature spaces, the linear SVM serves as the strongest classical baseline and primary non-transformer reference point.

To examine whether limited nonlinearity improves performance, we additionally evaluate Histogram Gradient Boosting (HGB). Because tree-based models are less effective on sparse high-dimensional TF–IDF inputs, dimensionality is first reduced using Truncated SVD with 50–150 components. HGB is then trained using multiclass log loss, with learning rate, iteration count, tree depth, and regularization tuned through validation and early stopping. While HGB can capture some nonlinear feature interactions, dimensionality reduction sacrifices fine-grained lexical structure, limiting overall gains.

3.2.2 Multilayer Perceptron (MLP)

Direct application of neural networks to sparse TF–IDF representations proved unstable in preliminary experiments, exhibiting elevated variance and minority-class collapse under smaller training subsets. To improve stability, we replace sparse lexical

inputs with dense sentence-level semantic embeddings generated using Sentence-BERT (SBERT), mapping each post to a fixed-length contextual vector.

We evaluate two pretrained SBERT encoders, MiniLM ("paraphrase-MiniLM-L6-v2" [1] with a total of 384 dimensions) and MPNet ("all-mpnet-base-v2" [2] with 768 dimensions), and consistently observe stronger performance with MPNet, indicating that embedding quality exerts greater influence than classifier complexity once stable representations are established. These embeddings serve as inputs to an MLP classifier for predicting candidate affiliation.

Our primary neural architecture consists of fully connected layers with batch normalization, GELU activations, dropout, and residual skip connections. Training uses multiclass cross-entropy loss with AdamW optimization, with early stopping based on validation macro-F1. To assess the necessity of architectural depth, we also evaluate a simplified single-hidden-layer MLP. Performance remains comparable, suggesting diminishing returns from deeper classifier architectures when fixed high-quality embeddings are used. Although full encoder fine-tuning could theoretically improve task specificity, empirical results show reduced stability and weaker generalization. Accordingly, we freeze the SBERT encoder and optimize only the MLP head, improving reproducibility, computational efficiency, and training stability.

3.2.3 Deep Learning

The final model family consists of BERTweet, a domain-adapted transformer pretrained on large-scale X data and specifically designed to capture platform-specific language patterns such as hashtags, slang, abbreviations, and informal syntax (Nguyen et al. 2020). Posts are tokenized using the BERTweet tokenizer and truncated or padded to a fixed sequence length. The contextualized representation of the `[CLS]` token is then passed through a dropout-regularized linear classification head to predict the three stance classes.

BERTweet is fine-tuned end-to-end using multiclass cross-entropy loss and AdamW optimization. This model is the highest-capacity text-only architecture in our study and serves as an empirical upper bound for contextual text modeling. Its domain-specific pretraining provides strong capacity for capturing implicit political cues, sarcasm, and nuanced rhetorical framing. However, while BERTweet achieves near-perfect performance on the full dataset, experiments on reduced training subsets show pronounced degradation, highlighting the sensitivity of transformer-based performance to dataset scale and reinforcing practical trade-offs between maximal predictive accuracy and deployment efficiency.

3.2.4 Multi-Feature Learning

Building on the feature representations described in Section 3.1.2 and the SBERT+MLP framework above, we investigate whether structured metadata improves candidate affiliation prediction beyond text-only representations. This analysis is conducted within the SBERT+MLP framework because it offers a strong balance among representational quality, computational tractability, and training stability.

[1] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2
[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Table 3: **Macro-F1** performance comparison across classical, neural, and transformer models under baseline and optimized configurations. Base models reflect initial implementations with minimal optimization, while optimal models incorporate full preprocessing, feature refinement, and hyperparameter tuning.

Model	**Log. Reg**	**SVM**	**HGB**	**MLP**	**BERTweet**
Base Models					
Test	0.683	0.754	0.643	0.683	0.643
Optimal Models					
Validation	0.922	0.955	0.840	0.852	0.996
Test	0.933	0.963	0.849	0.850	0.995

Non-textual features are grouped into four primary blocks: post-level numeric features, temporal indicators, geographic attributes, and platform-provided safety signals. These auxiliary features are concatenated with the 768-dimensional MPNet embedding and passed through the same MLP architecture used in the text-only setting. By holding the classifier structure constant, we isolate the marginal contribution of each metadata block and their combinations through systematic ablation analysis.

Results indicate that select features, particularly sentiment score and engagement rate, provide modest but consistent gains when introduced individually. However, combining multiple metadata categories frequently introduces redundancy, yielding limited improvement and, in some cases, degraded generalization. These findings suggest that while carefully selected behavioral and temporal features can complement textual representations, high-quality text embeddings remain the dominant predictive signal for political stance classification on X.

Overall, this analysis demonstrates that increased feature complexity does not inherently improve stance detection performance. Instead, principled feature selection and controlled integration are essential to avoid redundancy, noise amplification, and unnecessary modeling complexity in politically oriented OSN classification systems.

4 Results

This section presents empirical results for candidate affiliation prediction on X during the 2024 U.S. presidential election. Model performance is evaluated primarily using macro-averaged F1 to account for class imbalance across partisan and nonpartisan classes. Results are organized to compare performance across baseline text-only models, progressively refined preprocessing pipelines, learned semantic text representations, and structured multi-feature configurations. This staged analysis enables systematic evaluation of how preprocessing rigor, feature design, and model complexity jointly influence predictive accuracy, generalization, and practical trade-offs.

4.1 Evaluation of Models Using Text Embeddings

As described in Sections 3.2.1, 3.2.2, and 3.2.3, we first evaluate text-only stance detection performance across classical machine learning, sentence-embedding-based neural, and transformer-based architectures. This staged comparison isolates the effects of preprocessing rigor, representational quality, and architectural complexity without the influence of auxiliary metadata.

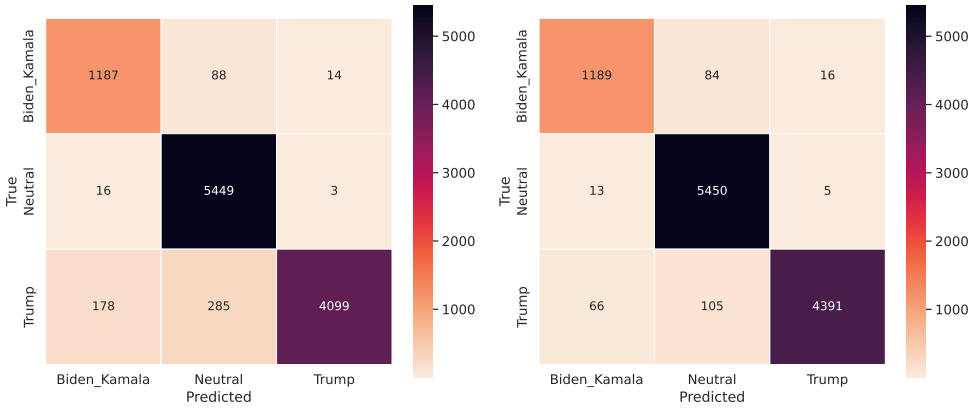
As shown in Table 3, baseline classical machine learning models trained on TF–IDF representations exhibit substantial performance variation. Among these approaches, the linear Support Vector Machine (SVM) achieves the strongest baseline performance, with a test macro-F1 of 0.75, outperforming Logistic Regression (0.68) and Histogram Gradient Boosting (HGB) (0.64). These findings reinforce prior evidence that linear classifiers remain highly effective for high-dimensional, sparse political text when lexical structure is preserved. After incorporating enhanced preprocessing, feature normalization, and hyperparameter optimization, both Logistic Regression and linear SVM demonstrate substantial gains, achieving test macro-F1 scores of 0.93 and 0.96, respectively. In contrast, HGB benefits less from optimization and continues to underperform relative to linear baselines, suggesting that compressed feature representations may reduce the fine-grained lexical fidelity required for this task.

Neural architectures built on sentence-level semantic embeddings provide a different performance profile. SBERT embeddings combined with MLP classifiers consistently outperform baseline TF–IDF models and achieve stable test macro-F1 scores near 0.85. Performance differences between shallow and deeper MLP variants remain minimal, indicating that embedding quality contributes more strongly to predictive performance than downstream classifier depth. These results suggest that once semantically rich sentence representations are established, increasing classifier complexity alone yields limited marginal benefit.

Among all evaluated text-only approaches, the domain-adapted transformer BERTweet achieves the strongest overall performance. After full preprocessing and end-to-end fine-tuning, BERTweet attains the highest macro-F1 on both the validation and test sets in this dataset. While this result highlights the effectiveness of domain-specific pretraining for modeling platform-specific language, implicit political cues, and contextual framing on X, it should be interpreted cautiously. Performance at this level likely reflects characteristics of the present dataset, including candidate-affiliation regularities, annotation structure, and task-specific lexical patterns, rather than implying near-perfect political stance detection across broader real-world conditions. More heterogeneous, ambiguous, or cross-domain political datasets would likely present substantially greater challenges. Accordingly, although transformer-based models demonstrate the highest predictive capacity under favorable in-domain conditions, these findings should be viewed primarily as evidence of strong dataset-specific performance rather than a general ceiling on stance detection difficulty. In contrast, carefully engineered classical models remain highly competitive while offering substantially lower computational cost and greater interpretability.

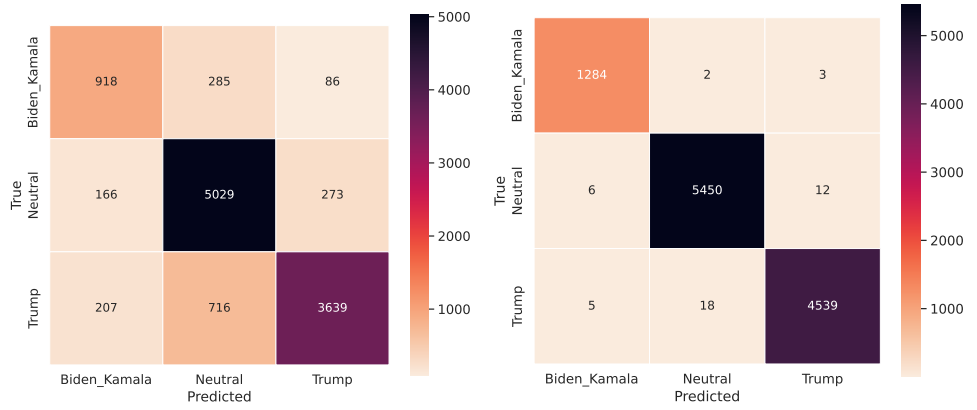
For class-level prediction performance, all models classify the Neutral class well, likely due to strong lexical or structural signals, as shown in Fig. 4. Misclassifications often collapse into the Neutral class, suggesting that Neutral functions as a “catch-all”

category, which is common in stance detection tasks. SVM and BERTweet show strong separation between Biden–Kamala and Trump. Most errors occur between partisan classes and Neutral rather than between opposing candidates. This error pattern is less harmful than false partisan attribution, but it may reduce sensitivity to real shifts in political stance.



(a) Logistic Regression

(b) SVM



(c) HGB

(d) BERTweet

Fig. 4: Confusion matrices for models on the test set, highlighting error patterns.

4.2 Multi-Feature ML Analysis

To assess the contribution of structured non-textual information, we conduct a controlled-feature-block analysis within the SBERT+MLP framework. Beginning with

Table 4: Performance across incremental multi-feature configurations within the SBERT+MLP framework. Feature blocks include post-only text (P), Content Semantics (A), Engagement Dynamics (B), Geo Context (C), and Temporal Behavior (D), added progressively to evaluate the marginal contribution of each feature category.

P	A	B	C	D	Validation		Test	
					Acc(%)	F1	Acc(%)	F1
✓	–	–	–	–	88.61	0.853	88.44	0.855
✓	✓	–	–	–	88.17	0.845	87.77	0.847
✓	✓	✓	–	–	88.27	0.848	88.05	0.849
✓	✓	✓	✓	–	87.70	0.840	88.21	0.851
✓	✓	✓	✓	✓	88.24	0.844	88.05	0.848

a text-only baseline (P), additional feature blocks are incrementally introduced following the ABCD structure described in Section 3. Table 4 summarizes validation and test performance across configurations.

The text-only baseline achieves strong performance, with a test accuracy of 88.44% and a macro-F1 of 0.855. Adding Block A alone slightly reduces performance (macro-F1 = 0.847), suggesting that auxiliary semantic features may introduce redundancy when combined with strong sentence embeddings. Incorporating Block B (engagement dynamics) partially recovers this decline, while adding Block C (geo-context) yields the highest intermediate test accuracy (88.21%), though gains remain modest. The full feature configuration (A+B+C+D) achieves balanced but not superior performance, with a test macro-F1 of 0.848, remaining broadly comparable to the text-only baseline.

Overall, these results indicate that high-quality text embeddings capture most of the predictive signal for candidate affiliation in this dataset. While selected metadata features can provide modest complementary value, broader feature expansion does not consistently improve generalization and may introduce redundancy or noise. These findings reinforce the idea that increased feature complexity alone does not guarantee improved performance in political stance detection.

4.3 Feature Ablation Study

To further isolate the contribution of individual metadata signals, we evaluate auxiliary features independently. Table 5 summarizes the performance impact of each feature when added separately to the baseline framework.

Across individual features, performance differences remain relatively modest, suggesting that no single auxiliary signal dramatically alters predictive capacity. Sentiment-related features yield the highest observed test macro-F1 (0.857), indicating that affective orientation may provide slightly stronger complementary value than the other evaluated features. Engagement rate and selected temporal indicators produce broadly comparable results, while posting hour and day of week show somewhat smaller gains. Given the narrow performance range across configurations,

Table 5: Independent performance contribution of individual auxiliary features within the SBERT+MLP framework. Validation and test metrics reflect the marginal predictive impact of each feature when evaluated separately relative to the baseline text representation.

	Validation		Test	
Feature	**Acc (%)**	**F1**	**Acc (%)**	**F1**
engagement_rate_log	88.67	0.852	88.20	0.852
tweet_dayofweek	88.53	0.852	88.17	0.852
sentiment_score	88.49	0.850	88.48	0.857
tweet_hour	88.30	0.848	88.13	0.852

these findings should be interpreted primarily as incremental trends rather than large performance differentials.

Overall, the results suggest that individual auxiliary features can provide modest complementary information, but their contributions remain secondary to the underlying text representation. High-quality linguistic and embedding-based features continue to drive the majority of predictive performance, while selective metadata integration may offer only a limited incremental benefit in this dataset and evaluation setting.

5 Discussion

We interpret the empirical findings in the context of prior work in computational social science (CSS) and the practical demands of political discourse analysis on online social networks. Rather than emphasizing predictive performance alone, we focus on how preprocessing rigor, feature representation, model complexity, and deployment constraints jointly shape robustness, interpretability, and practical utility. Across experiments, the results reveal consistent trade-offs: while higher-capacity models can achieve stronger in-domain performance under favorable conditions, carefully engineered classical and hybrid approaches often remain highly competitive while offering advantages in transparency, efficiency, and reproducibility. These findings reinforce that scalable political stance detection depends not solely on maximizing predictive scores, but on balancing accuracy with methodological accountability and deployment feasibility.

5.1 Trade-offs in Political Stance Detection

Across all modeling paradigms, the results suggest that feature representation quality and preprocessing rigor exert a greater influence on performance than architectural complexity alone. Although transformer-based models achieve the strongest results within this dataset, the substantial gains observed in optimized classical models indicate that principled preprocessing, lexical normalization, and task-aligned feature design can dramatically narrow the apparent gap between simpler and more complex architectures. This finding aligns with the broader CSS literature, which emphasizes

that strong data preparation and representation remain foundational to robust social media analysis.

The strong performance of linear SVMs is particularly notable. Political discourse on X frequently contains recurring lexical structures, including candidate references, campaign slogans, partisan phrases, and issue framing, which can often be effectively captured through sparse high-dimensional representations when noise is sufficiently reduced. Under these conditions, linear models offer important practical advantages beyond predictive performance, including interpretability, computational efficiency, reproducibility, and lower deployment cost. For large-scale monitoring systems or policy-relevant applications where transparency and auditability are essential, such properties may be as important as marginal predictive improvements.

Sentence-embedding-based neural models provide a useful middle ground by improving semantic representation while maintaining substantially lower complexity than full transformer fine-tuning. However, our results also indicate diminishing returns from increasing classifier depth once high-quality embeddings are established, suggesting that representational quality contributes more strongly than downstream architectural complexity. This observation supports a broader principle emerging across the experiments: stronger representations do not necessarily require maximal model complexity.

The feature ablation and multi-feature analyses further reinforce this pattern. Auxiliary metadata, including sentiment, engagement, temporal, and geographic signals, provides only modest and often inconsistent gains relative to strong text representations. Given the relatively narrow performance differences across many feature configurations, these effects should be interpreted cautiously. While selected metadata may offer incremental complementary value, broader feature expansion does not consistently improve generalization and may introduce redundancy, noise, or additional system complexity. Collectively, these findings caution against assuming that additional features inherently improve performance in political stance detection.

5.2 Implications for Election Monitoring and Responsible Deployment

From an applied perspective, these findings suggest that effective election-monitoring systems should prioritize transparency, robustness, and computational practicality alongside predictive accuracy. While transformer-based systems demonstrate strong in-domain predictive capacity, their computational cost, reduced interpretability, and sensitivity to dataset structure may complicate deployment in real-time, policy-sensitive, or resource-constrained environments. In contrast, classical and hybrid neural systems may offer more favorable operational trade-offs when scalability, auditability, and methodological clarity are primary objectives.

Importantly, the near-perfect BERTweet performance observed in this study should be interpreted cautiously. As discussed in the Results section, performance at this level likely reflects favorable characteristics of the present dataset, including candidate-affiliation regularities, annotation structure, and domain-specific lexical patterns, rather than implying that political stance detection has been broadly solved. More heterogeneous datasets, adversarial political discourse, multilingual settings,

or cross-platform deployment would likely present substantially greater challenges. Accordingly, transformer performance here should be viewed primarily as a strong dataset-specific capacity rather than a universal benchmark for real-world stance detection.

It is also important to distinguish candidate affiliation labeling from broader political behavior. The labels in this study capture candidate-associated discourse patterns expressed on X, not verified political ideology, voting behavior, or long-term partisan identity. As such, these models are best understood as tools for analyzing online communicative alignment within a defined platform and temporal context rather than direct measures of voter intent or political belief.

Overall, the results emphasize that reliable political stance detection is less a function of maximal complexity than of principled design choices: rigorous preprocessing, careful feature selection, transparent annotation procedures, and controlled comparative evaluation. These considerations are particularly important in politically sensitive applications, where methodological overreach may lead to misleading claims.

5.3 Limitations & Future Work

Several limitations should temper interpretation of these findings. First, this study is restricted to English-language X data from a single U.S. election cycle, limiting broader generalizability across platforms, political systems, languages, and evolving discourse environments. Second, although candidate affiliation labeling demonstrated strong inter-rater reliability, heuristic and keyword-guided annotation cannot fully eliminate ambiguity, sarcasm, or context-dependent stance uncertainty. Third, model performance may reflect dataset-specific regularities that may not transfer robustly under distribution shift, especially as political discourse evolves over time.

Future work should therefore prioritize broader robustness evaluation rather than solely pursuing higher benchmark performance. Important directions include cross-platform analysis, multilingual stance detection, domain adaptation across election cycles, and evaluation under more ambiguous or adversarial discourse settings. Hierarchical classification frameworks that explicitly separate neutral detection from partisan classification may also improve interpretability. Additional work on sarcasm detection, temporal adaptation, confidence calibration, and post-hoc uncertainty estimation could further strengthen practical deployment reliability.

Ultimately, the broader contribution of this work lies less in establishing a single best-performing model than in clarifying how preprocessing, representation, feature complexity, and model choice interact under realistic conditions of political discourse. Continued emphasis on reproducibility, cautious interpretation, and deployment-aware evaluation will be essential for developing trustworthy social media analysis systems in future politically sensitive contexts.

6 Conclusion

Accurate political stance detection on online social networks is important for analyzing political discourse, but reliable deployment requires more than maximizing predictive accuracy. This study examined how preprocessing, feature representation, and model

selection influence candidate affiliation detection on X during the 2024 U.S. presidential election through a unified comparison of classical, neural, and transformer-based models.

The results show that rigorous preprocessing and task-aligned text representations often matter more than architectural complexity alone. While transformer-based models achieve the strongest in-domain performance, these gains likely reflect dataset-specific regularities. Classical models, particularly linear SVMs, remain highly competitive while offering advantages in efficiency, interpretability, and reproducibility. Multi-feature analyses further indicate that expanded metadata provides only limited gains beyond strong text representations.

Overall, robust political stance detection depends more on principled methodological design than maximal model complexity. These findings offer practical guidance for scalable and trustworthy political discourse analysis.

Data Availability. Data and analysis scripts associated with this manuscript are publicly available at the following link: https://github.com/Sensify-Lab/election_detection

Statements and Declarations. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. No funding was received to assist with the preparation of this manuscript.

References

A. Semary N, Ahmed W, Amin K, et al (2024) Enhancing machine learning-based sentiment analysis through feature extraction techniques. Plos one 19(2):e0294968. https://doi.org/10.1371/journal.pone.0294968

Aamir M, J L, Sweety (2024) A comparative study of ml and dl approaches for twitter sentiment classification. In: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), pp 1–5, https://doi.org/10.1109/ICEECT61758.2024.10738938

Albladi A, Islam M, Seals C (2025) Sentiment analysis of twitter data using nlp models: a comprehensive review. IEEE Access https://doi.org/10.1109/access.2025.3541494

Aldayel A, Magdy W (2019) Your stance is exposed! analysing possible factors for stance detection on social media. Proceedings of the ACM on human-computer interaction 3(CSCW):1–20. https://doi.org/10.1145/3359307

Alturayeif N, Luqman H, Ahmed M (2023) A systematic review of machine learning techniques for stance detection and its applications. Neural Computing and Applications 35(7):5113–5144. https://doi.org/10.1007/s00521-023-08285-7

Alvi Q, Ali SF, Ahmed SB, et al (2023) On the frontiers of twitter data and sentiment analysis in election prediction: a review. PeerJ Computer Science 9:e1517. https://doi.org/10.7717/peerj-cs.1517

Boutet A, Kim H, Yoneki E (2012) What's in your tweets? i know who you supported in the uk 2010 general election. In: Proceedings of the international AAAI conference on web and social media, pp 411–414, https://doi.org/10.1609/icwsm.v6i1.14283

Brena G, Brambilla M, Ceri S, et al (2019) News sharing user behaviour on twitter: A comprehensive data collection of news articles and social interactions. In: Proceedings of the International AAAI Conference on Web and Social Media, pp 592–597, https://doi.org/10.1609/icwsm.v13i01.3256

Chaudhry HN, Javed Y, Kulsoom F, et al (2021) Sentiment analysis of before and after elections: Twitter data of us election 2020. Electronics 10(17):2082. https://doi.org/10.3390/electronics10172082

Chauhan DS, Kumar R, Ekbal A (2019) Attention based shared representation for multi-task stance detection and sentiment analysis. In: International conference on neural information processing, Springer, pp 661–669, https://doi.org/10.1007/978-3-030-36802-9_70

Conover M, Ratkiewicz J, Francisco M, et al (2011) Political polarization on twitter. In: Proceedings of the international aaai conference on web and social media, pp 89–96, https://doi.org/10.1609/icwsm.v5i1.14126

Darwish K, Stefanov P, Aupetit M, et al (2020) Unsupervised user stance detection on twitter. In: Proceedings of the international AAAI conference on web and social media, pp 141–152, https://doi.org/10.1609/icwsm.v14i1.7286

Haouari F, Elsayed T (2024) Are authorities denying or supporting? detecting stance of authorities towards rumors in twitter. Social Network Analysis and Mining 14(1):34. https://doi.org/https://doi.org/10.1007/s13278-023-01189-3

Jonnala NS, Ram Teja A, Rajeswari SR, et al (2025) Leveraging hybrid model for accurate sentiment analysis of twitter data. Scientific Reports 15(1):24438. https://doi.org/10.1038/s41598-025-09794-2

Kejriwal M, Wang Q, Li H, et al (2021) An empirical study of emoji usage on twitter in linguistic and national contexts. Online social networks and media 24:100149. https://doi.org/10.1016/j.osnem.2021.100149

Kowsik VS, Yashwanth L, Harish S, et al (2024) Sentiment analysis of twitter data to detect and predict political leniency using natural language processing. Journal of Intelligent Information Systems 62(3):765–785. https://doi.org/10.1007/s10844-024-00842-3

Küçük D, Can F (2020) Stance detection: A survey. ACM Computing Surveys (CSUR) 53(1):1–37. https://doi.org/10.1145/3369026

Küçük D, Can F (2022) A tutorial on stance detection. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp 1626–1628, https://doi.org/10.1145/3488560.3501391

Kumari S, Singh MP (2024) Machine learning-based election results prediction using twitter activity. SN Computer Science 5(7):819. https://doi.org/10.1007/s42979-024-03180-x

Mansur Z, Omar N, Tiun S (2023) Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. IEEE Access 11:16226–16249. https://doi.org/10.1109/access.2023.3239375

Marshan A, Nizar FNM, Ioannou A, et al (2025) Comparing machine learning and deep learning techniques for text analytics: Detecting the severity of hate comments online. Information Systems Frontiers 27(2):487–505. https://doi.org/10.1007/s10796-023-10446-x

McHugh ML (2012) Interrater reliability: the kappa statistic. Biochemia medica 22(3):276–282. https://doi.org/10.11613/bm.2012.031

Mohammed Alsekait D, Fathi H, Abdallah Ibrahim S, et al (2025) Sentiment analysis: A machine learning utilisation for analyzing the sentiments of facebook and twitter posts. Intelligent Data Analysis 29(4):889–912. https://doi.org/10.1177/1088467x241301389

Mullins LE, Malkin A, Riosa PB, et al (2025) # contentanalysisofabaontwitter/x: Finding behavior analysis' heart on social media. Behavior Analysis in Practice pp 1–15. https://doi.org/10.1007/s40617-025-01065-w

Nguyen DQ, Vu T, Nguyen AT (2020) Bertweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 9–14, https://doi.org/10.18653/v1/2020.emnlp-demos.2

Padalko H, Chumachenko D, Kaur N, et al (2025) Investigation of machine learning approaches to classify war-related content during russian full-scale invasion of ukraine. Social Network Analysis and Mining 15(1):73. https://doi.org/https://doi.org/10.1007/s13278-025-01501-3

Pandya V, Somthankar A, Shrivastava SS, et al (2021) Twitter sentiment analysis using machine learning and deep learning techniques. In: 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), IEEE, pp 1–5, https://doi.org/10.1109/c2i454156.2021.9689241

Papadimitriou O, Kanavos A, Maragoudakis M, et al (2024) A comparative study of sentiment analysis on twitter and reddit using deep learning techniques. In: Proceedings of the 2024 8th International Conference on Algorithms, Computing and Systems, pp 50–55, https://doi.org/10.1145/3708597.3708605

Patel S, Bansal P, Kaur P (2024) Rumour detection on benchmark twitter datasets using graph neural networks with data augmentation. Social Network Analysis and Mining 14(1):163. https://doi.org/https://doi.org/10.1007/s13278-024-01328-4

Qi Y, Shabrina Z (2023) Sentiment analysis using twitter data: a comparative application of lexicon-and machine-learning-based approach. Social network analysis and mining 13(1):31. https://doi.org/10.1007/s13278-023-01030-x

Rustam F, Khalid M, Aslam W, et al (2021) A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. Plos one 16(2):e0245909. https://doi.org/10.1371/journal.pone.0245909

Skoularikis K, Savvas I, Garani G, et al (2026) Hybrid sentiment analysis approach with domain-specific lexicons with active learning on greek social media texts. Social Network Analysis and Mining https://doi.org/https://doi.org/10.1007/s13278-025-01560-6

Wang Q, Erqsous M, Barner KE, et al (2025) Lata: A pilot study on llm-assisted thematic analysis of online social network data generation experiences. Proceedings of the ACM on Human-Computer Interaction 9(2):1–28. https://doi.org/10.1145/3711022

Wang Q, Khatiwada P, Vieira CC, et al (2026) Wisdom of the llm crowd: A large scale benchmark of multi-label u.s. election-related harmful social media content. arXiv:2602.11962

Zhang X (2025) Sentiment and stance classification on climate change tweets using fine-tuned bert models: A domain-adaptive nlp framework for analyzing public discourse in social media. In: Proceedings of the 2025 7th International Conference on Big Data Engineering, pp 31–36, https://doi.org/10.1145/3775050.3775055

Zubiaga A, Kochkina E, Liakata M, et al (2018) Discourse-aware rumour stance classification in social media using sequential classifiers. Information Processing & Management 54(2):273–290. https://doi.org/10.1016/j.ipm.2017.11.009