Understanding the Linguistic Cues Behind Stance Detection

Parush Gera
University of South Florida
parush@usf.edu

Tempestt Neal
University of South Florida
tjneal@usf.edu

Abstract

Stance detection seeks to determine whether a text expresses a position in favor of, against, or neutral toward a target. Despite advances in neural architectures, performance remains inconsistent across datasets. To better understand these disparities, we analyze over 75K samples from four benchmark datasets using six neural models, focusing on stylistic and pragmatic language features rather than architectures or external knowledge. We extract 43 features spanning lexical richness, syntactic complexity, affective tone, and hedging, and assess their impact through both Logistic Regression and SHAP analyses. Our findings reveal distinct stylistic profiles for each stance: favor is best detected when expressed concisely with minimal hedging; against when paired with strong negative emotions and greater lexical variety; and none when texts are lexically simple and emotionally neutral. Across classes, errors arise from excessive complexity, mixed emotional signals, and overuse of hedging. These results advance understanding of what drives success and failure in stance detection.

1 Introduction

Stance detection is a natural language opinion mining task that determines whether a text expresses a position *in favor of*, *against*, or *neutral/none* toward a specific target, such as an organization, individual, policy, or other identifiable entity (Gera and Neal, 2025; Küçük and Can, 2020; Mohammad et al., 2016). It underpins applications in public opinion tracking (Glandt et al., 2021), misinformation detection (Hardalov et al., 2022), political discourse analysis (Li et al., 2021), and studies of social dynamics in online communities (Allaway and McKeown, 2020). Advances in deep learning have driven notable gains, with recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based architectures such as BERT capturing increasingly rich contextual representations (Ghosh et al., 2019; Xu et al., 2018; Augenstein et al., 2016; Mohammad et al., 2016). Nonetheless, performance remains highly variable across datasets and models: recent studies report F_1-scores ranging from 0.43 for Gated Recurrent Units to 0.87 for BERT (Devlin et al., 2019), even as large-scale language models (LLMs) and generative variants enter the field (Cruickshank and Ng, 2025). This variability raises a central question: *which linguistic and contextual signals do models rely on when detecting stance, and among these, which consistently drive performance gains versus those that fluctuate across settings?*

Unlike sentiment analysis, which classifies text by polarity, or aspect-based sentiment analysis (ABSA), which links opinions to entity attributes (Chaudhary and Neal, 2026), stance detection centers on the author's position toward a specific target that may be implicit, requiring inference from context (Allaway and McKeown, 2020). This implicit nature of stance expression—where a text's position toward a target may be unstated, indirect, or embedded in framing rather than overt sentiment—has two important consequences. First, it suggests that stance detection depends on linguistic and contextual cues beyond the explicit polarity markers that sentiment analysis often relies on. Yet the identity of these stance-specific signals remains unclear, as performance varies widely across models (Gera and Neal, 2025), indicating that different architectures capture them inconsistently. Second, implicit stance pushes models beyond surface semantics, requiring world knowledge, target-aware attention, and contextual reasoning (He et al., 2022). Recent approaches augment pretrained language models with knowledge graphs, contrastive learning, or multi-task objectives (Liang et al., 2022a; He et al., 2022), yet robustness studies reveal steep performance drops under domain shift, figurative language, or dialec-

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 302–316

tal variation (Calderon et al., 2024; Jhamtani et al., 2021). We hypothesize that this fragility stems not only from topical differences but also from shifts in stylistic and pragmatic cues that shape how stance is expressed.

Prior work has primarily emphasized improving model architectures or incorporating external knowledge (Gera and Neal, 2025), while giving less attention to stylistic and pragmatic cues. Hand-engineered stylistic features capturing syntactic, lexical, and semantic complexity have been widely examined in related NLP tasks such as sentiment analysis, deception detection, and fake news classification (Tan et al., 2023; Pekar et al., 2024; Fahmy et al., 2025; Yang et al., 2024), where they have improved interpretability, revealed sources of model error, and highlighted linguistic cues that reliably signal meaning or intent. We hypothesize that similar cues shape how stance is expressed and detected. Guided by this motivation, our study addresses two core research questions:

1. Which stylistic and pragmatic features distinguish texts that stance detection models consistently classify correctly from those they misclassify?
2. What do these features reveal about the linguistic strategies authors use to express stance, and how are such expressions interpreted by neural models?

To address these questions, we trained six diverse stance detection models—two GRU variants, two LSTM variants, a CNN, and BERT—on a unified corpus of four benchmark datasets. We identified test examples that all models classified consistently, either correctly or incorrectly, yielding two stable sets for comparison. From each example, we extracted 43 linguistically grounded stylistic features spanning lexical richness, syntactic complexity, and affective tone. To isolate the most predictive signals, we adopted a twofold analytical framework: Logistic Regression, to capture linear relationships between individual stylistic features and classification outcomes, and SHAP (SHapley Additive exPlanations), to provide model-agnostic insights into non-linear interactions among features and their combined influence on those outcomes. This integrated approach enables us not only to rank features by the consensus of both methods but also to analyze how specific feature values contribute to correct or incorrect predictions, thereby uncovering the linguistic patterns that shape robust stance detection[1].

2 Datasets

We use four publicly available stance detection datasets, summarized in Table 1:

- **SemEval-2016 Task 6** (Mohammad et al., 2016): Tweets manually annotated as *favor*, *against*, or *none* toward six sociopolitical targets.
- **PStance** (Li et al., 2021): Tweets labeled *favor* or *against* toward three U.S. political figures, collected around the 2020 presidential election and annotated using weak supervision.
- **COVID-19 Stance** (Glandt et al., 2021): Tweets manually labeled as *favor*, *against*, or *none* toward four public health-related targets.
- **WT-WT** (Conforti et al., 2020): The largest publicly available English stance detection dataset, consisting of tweets about mergers and acquisitions labeled *support*, *refute*, *comment*, or *unrelated*. We mapped *support* to *favor*, *refute* to *against*, and both *comment* and *unrelated* to *none*.

Dataset	Target or Domain	Train	Test
COVID-19	face_masks	1,507	200
	fauci	1,664	200
	school_closures	990	200
	stay_at_home_orders	1,172	200
	Total	**5,333**	**800**
PStance	Bernie Sanders	5,690	635
	Donald Trump	7,176	777
	Joe Biden	6,551	745
	Total	**19,417**	**2,157**
SemEval-2016	Atheism	513	220
	Climate Change is a Concern	395	169
	Donald Trump	530	177
	Feminist Movement	664	285
	Hillary Clinton	689	295
	Legalization of Abortion	653	280
	Total	**3,444**	**1,426**
WT-WT	Healthcare Domain	22,101	7,367
	Entertainment Domain	11,141	3,714
	Total	**33,242**	**11,081**
Combined	—	**61,436**	**15,646**

Table 1: Dataset statistics with train/test splits per target.

We unified all four datasets into a single combined corpus for training (>60,000 samples) and evaluation (>15,000 samples). These datasets collectively cover a wide spectrum of stance targets—including public figures, political policies, health

[1]Code and configurations are available at https://github.com/parushgera/linguistic-cues-stance-detection.

topics, and entertainment speculation—and contain rich variability in linguistic style, tone, and rhetorical strategies. This unified framework ensured that the stylistic patterns we later analyzed reflected generalizable linguistic correlates of stance rather than domain-specific idiosyncrasies, and enabled us to identify examples robustly classified across architectures.

3 Models

We trained six neural models representing the major architecture families used in stance detection (Liang et al., 2022b; He et al., 2022; Xu et al., 2018; Allaway and McKeown, 2020; Zhang et al., 2020a): **BiGRU** (Chung et al., 2014) and **Att-BiGRU** (Zhou et al., 2017) (bidirectional GRUs, with and without attention); **BiLSTM** (Schuster and Paliwal, 1997) and **Att-BiLSTM** (Siddiqua et al., 2019) (bidirectional LSTMs, with and without attention); **KimCNN** (Kim, 2014) (a CNN applying multiple filter widths over embeddings); and **BERT** (Devlin et al., 2019), fine-tuned with a classification head over the [CLS] token.

For all non-transformer models, we used contextualized token-level embeddings from the final hidden layer of the `all-MiniLM-L6-v2` Sentence-BERT (SBERT) model (Reimers and Gurevych, 2019), yielding 384-dimensional representations per token. Unlike static embeddings such as GloVe (Pennington et al., 2014) used in earlier stance detection work (Augenstein et al., 2016; Xu et al., 2018; Du et al., 2017), SBERT embeddings encode contextual semantics, which is especially valuable for user-generated content where lexical meaning shifts depending on context. Each sequence was padded or truncated to 128 tokens. For BERT, we used its own contextualized representations learned during fine-tuning.

All models were trained using the Optuna hyperparameter optimization framework (Akiba et al., 2019), with macro-F_1 as the objective metric. Each model underwent 100 optimization trials; the search space was tailored per architecture (see Appendix). Unlike many stance detection studies that focus on in-target or cross-target settings, our experiments adopt a multi-target setting: models were trained on a combined dataset of 61,436 examples spanning multiple targets and evaluated on a held-out test set of 15,646 examples. This design provides a broader view of how stylistic features operate across diverse targets rather than being restricted to a single domain.

4 Methodology

Following training, each model was evaluated on a held-out test set ($n = 15{,}646$). Table 2 summarizes the macro-F_1 scores achieved by each model. BERT achieved the highest F_1 score despite SBERT embeddings providing contextual representations to the other models. This gap likely reflects BERT's end-to-end fine-tuning, which adapts its internal representations to the stance detection task, whereas non-transformer models use SBERT embeddings as fixed inputs. To isolate stylistic cues that consistently influence stance detection performance, rather than artifacts of individual architectures, we restricted our analysis to examples that were either uniformly classified correctly or uniformly misclassified by all six models. This procedure yielded 9,769 consistently correct and 758 consistently misclassified examples, totaling 10,527 instances (≈67% of the test set). The remaining ≈33% of test instances, where models disagreed, were excluded to ensure that the stylistic patterns we identified reflected consistent architectural behavior rather than model-specific variation. While these ambiguous cases are themselves of interest, analyzing them would conflate architecture-specific biases with linguistic effects, which we leave to future work.

Table 2: Macro-F_1 scores for stance detection models on the test set.

Model	F_1 Score
BiGRU	0.7803
Att-BiGRU	0.7936
BiLSTM	0.8006
Att-BiLSTM	0.7782
KimCNN	0.7937
BERT	0.8258

4.1 Feature Extraction

From all 10,527 consensus samples, we extracted 43 hand-crafted features spanning six categories (Table 3):

- **Lexical Richness:** Word usage and vocabulary diversity (e.g., type-token ratio, hapax legomena, stopword ratio).
- **Part-of-Speech Composition:** Proportions of grammatical categories using Penn-Treebank tags (Bird and Loper, 2004).

- **Syntactic and Structural Complexity:** Sentence structure indicators (e.g., function word ratio, subordinate clause count, Flesch–Kincaid readability). We note that punctuation count and Flesch–Kincaid readability are included here as they capture surface-level structural properties (sentence segmentation and sentence-length-based complexity, respectively), though they are not syntactic features in a strict grammatical sense.
- **Affective Tone:** Sentiment polarity and subjectivity via TextBlob (Loria, 2018), and fine-grained emotion scores (anger, joy, fear, sadness, disgust, surprise) from a DistilRoBERTa model fine-tuned on diverse emotion-labeled datasets (Hartmann, 2022).
- **Epistemic Stance and Hedging:** Certainty and doubt markers from the MPQA lexicon (Wilson et al., 2005); sentence-level hedging scores from BERTweet-Hedge; and token-level hedge categories (C: certain, D: doxatic, E: epistemic, I: investigation, N: condition) from the hedgehog NER classifier.
- **Derived Uncertainty Ratios:** Hedging category counts normalized by total hedge words, yielding proportional uncertainty measures.

Where applicable, all count-based features were normalized by total word count to ensure comparability across texts of varying lengths.

Table 3: Overview of the six stylistic feature categories and the specific features extracted in each.

Feature Category	Features
Lexical Richness	(1) Total word count, (2) Sentence count, (3) Avg. word length, (4) Std. word length, (5) Type-token ratio, (6) Hapax legomena count, (7) Stopword ratio, (8) Punctuation density
Part-of-Speech	(9) Noun ratio, (10) Verb ratio, (11) Adjective ratio, (12) Adverb ratio, (13) Pronoun ratio
Syntactic/Structural	(14) Function word ratio, (15) Subordinate clause count, (16) Punctuation counts, (17) Sentence length variability, (18) Flesch–Kincaid readability
Affective Tone	(19) Sentiment polarity, (20) Subjectivity, (21) Anger, (22) Joy, (23) Fear, (24) Sadness, (25) Disgust, (26) Surprise
Epistemic Stance & Hedging	(27–29) Certainty adverbs/verbs/adjectives count, (30–32) Doubt adverbs/verbs/adjectives count, (33) Hedges score, (34–38) Hedgehog C/D/E/I/N count
Uncertainty Ratios	(39) C_ratio, (40) D_ratio, (41) E_ratio, (42) I_ratio, (43) N_ratio

4.2 Feature Analysis

4.2.1 Validation of Hand-Engineered Features

To validate that our hand-engineered features carry predictive signal, we trained Logistic Regression (LR) and XGBoost classifiers (Chen and Guestrin, 2016) to distinguish between texts that all six stance models classified correctly versus those that all models misclassified. LR captures direct, interpretable relationships between features and outcomes, while XGBoost models complex, non-linear feature interactions, offering a complementary perspective.

Because of substantial class imbalance (9,769 correct vs. 758 misclassified), we generated 13 balanced subsets ($\lceil 9{,}769/758 \rceil = 13$), each pairing all 758 misclassified examples with a distinct random sample of 758 correctly classified examples (without replacement), ensuring that every correctly classified instance was represented across the full set of iterations. To control for stance category, we applied this sampling independently for each label (*favor*, *against*, *none*), ensuring that correctly and incorrectly classified texts were compared within the same stance rather than across different stances.

Both classifiers achieved moderate predictive performance: for *favor*, LR reached $F_1 = 0.68 \pm 0.10$ and XGBoost 0.65 ± 0.08; for *against*, LR 0.59 ± 0.11 and XGBoost 0.59 ± 0.10; for *none*, LR 0.77 ± 0.05 and XGBoost 0.79 ± 0.06. The close correspondence between models confirms that the predictive signal is not an artifact of a particular classifier.

4.2.2 Feature Attribution

For feature attribution, we employed two complementary approaches: LR coefficients, which indicate the direction and magnitude of each feature's linear contribution, and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017), a model-agnostic framework that quantifies each feature's contribution by approximating Shapley values from cooperative game theory. Unlike regression coefficients, which assume a fixed linear effect, SHAP values are computed from the model's prediction function itself, allowing them to capture non-linear effects and interactions where a feature's contribution can vary depending on the values of other features. We computed SHAP values from LR models rather than XGBoost to maintain interpretive consistency: using the same underlying model for both coefficient and SHAP analysis ensures that any divergence between the two rankings reflects methodological differences (linear weights vs. marginal contributions) rather than differences in the underlying classifier. We analyzed features along two

dimensions: *magnitude* (overall strength of association with classification outcomes, addressing **RQ1**) and *direction* (whether higher or lower feature values promote correct vs. incorrect classification, addressing **RQ2**).

Magnitude Analysis. We aggregated mean absolute LR coefficients and mean absolute SHAP values across all 13 iterations, producing two complete rank-ordered lists of the 43 features. To ensure reliability, we combined these rankings through consensus-based quartile analysis: each list was divided into four quartiles (Q1: ranks 1–11, Q2: 12–22, Q3: 23–33, Q4: 34–43), and a feature was assigned to a quartile only if its position in *both* rankings fell within the same tier. This procedure ensured that quartile assignments reflected stable importance across both linear (coefficient-based) and model-agnostic (SHAP-based) perspectives, reducing the influence of any single method.

Direction Analysis. A feature's high magnitude does not imply that its effect is uniform; its influence may depend on context. To capture this, we examined the direction of each feature's SHAP contributions. A positive SHAP value indicates a feature pushed the prediction toward *correct*, while a negative value indicates a push toward *incorrect* classification. For each feature, we segregated instances into these two groups across the 13 iterations and calculated the mean feature value within each group. This reveals the specific feature value ranges associated with model success or failure—for example, whether high anger values are linked to correct classifications while lower values are associated with errors. All features were standardized (zero mean, unit variance) prior to analysis, ensuring that mean values reported in Tables 4–6 are directly comparable across features with different natural scales.

5 Results

We report findings separately for *favor*, *against*, and *none*, focusing on the most influential (Q1) features identified by our consensus quartile analysis. Full quartile-level breakdowns are provided in the Appendix (Tables 8–10). We note that some Q1 features exhibit near-equal frequency splits (e.g., ≈50/50), indicating that their directional influence was not strongly consistent across instances. These features were nonetheless retained in Q1 because their *magnitude*—the overall strength of their association with classification outcomes—was consistently high across both LR and SHAP rankings. Frequency indicates consistency of direction, while quartile assignment reflects overall importance.

5.1 *Favor* Predictions

Table 4 presents the Q1 features for *favor*, with rank agreement between LR and SHAP shown in Figure 1 ($\rho = 0.852$). A prominent pattern involves epistemic markers: doubt-related adverbs (e.g., "perhaps," "maybe") appeared more often in misclassified cases (+0.918, 26.8%) than in correct ones (−0.445, 73.2%), and certainty verbs (e.g., "prove," "know") showed a similar tendency, with higher values coinciding with errors (+0.477, 33.5%). This suggests that when support is either hedged or expressed with strong certainty, models find it harder to interpret the stance.

Part-of-speech features also played a role: correct predictions were more common in texts with lower noun ratios (−0.656, 59.7%) and lower pronoun ratios (−0.248, 55.3%). The remaining Q1 features captured lexical richness, generally indicating that more complex texts are harder to classify—shorter texts with lower word counts and fewer sentences were associated with correct predictions, while longer, more punctuated texts were more often misclassified.

Taken together, models perform best when support is expressed in simple, direct language—without hedging or excessive structural complexity. When stances are embedded in longer or epistemically nuanced text, models struggle to detect support reliably, suggesting a reliance on surface-level cues.

Table 4: Directional influence of Q1 features for *favor*. **Mean**: average standardized feature value when SHAP contribution was positive (correct) or negative (incorrect) across 13 iterations. **Freq.**: proportion of instances with positive/negative contribution.

	Correct		Incorrect		Feature Category					
Q1 Features	**Mean**	**Freq.**	**Mean**	**Freq.**	**Lex.**	**PoS**	**Syn.**	**Aff.**	**Hdg.**	**Unc.**
Doubt Adverbs Count	−0.445	73.2%	0.918	26.8%					X	
Certainty Verbs Count	−0.241	66.5%	0.477	33.5%					X	
Noun Ratio	−0.656	59.7%	0.647	40.3%		X				
Pronoun Ratio	−0.248	55.3%	0.013	44.7%		X				
Sentence Count	−0.036	53.3%	0.320	46.7%	X					
Punctuation Density	−0.113	52.2%	0.248	47.8%	X					
Total Word Count	−0.317	50.7%	0.364	49.3%	X					
Hapax Legomena	−0.131	50.5%	−0.006	49.5%	X					
Stopword Ratio	−0.645	49.4%	0.518	50.6%	X					
Type-Token Ratio	0.010	49.0%	−0.135	51.0%	X					
Total by Category					**6**	**2**	**0**	**0**	**2**	**0**

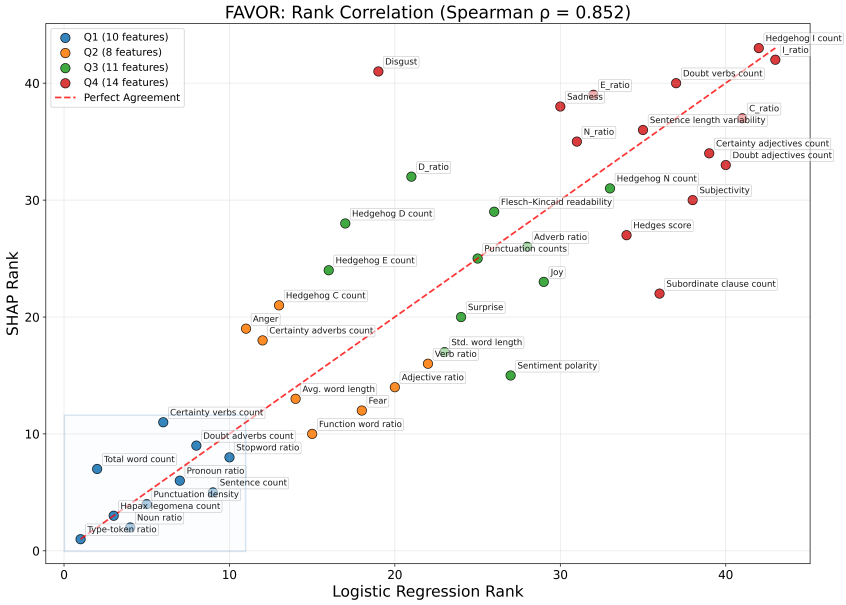

Figure 1: Rank correlation between LR coefficients and SHAP importance for *favor* ($\rho = 0.852$).

5.2 *Against* Predictions

Table 5 shows that Q1 features for *against* span lexical richness, part-of-speech, syntactic structure, and affect—contrasting with *favor*, where no Q1 features were drawn from syntax or affect. Rank agreement was moderate (Figure 2; $\rho = 0.676$).

Affective signals emerged as influential cues: stronger fear (+0.580, 52.6%) and anger (+1.196, 31.3%) were associated with correct classifications, while lower values for both coincided with misclassifications. Among structural features, higher punctuation density favored correct predictions (+0.170, 53.9%), while lower function word ratios (−0.915, 53.9%) and verb ratios (−0.705, 53.6%) also supported correct classification, indicating that models perform better when opposition is less grammatically dense.

As with *favor*, lower noun ratios and fewer sentences were associated with correct predictions. However, unlike *favor*, higher type–token ratios were linked to correct classifications (+0.561, 49.7%), suggesting that lexical variety supports detection of opposition when combined with emotional directness.

These results indicate that models rely most strongly on direct, emotionally charged expressions of opposition. When opposition is expressed in a neutral, formal, or structurally complex style, models struggle to classify it accurately.

Table 5: Directional influence of Q1 features for *against*. **Mean**: average standardized feature value when SHAP contribution was positive (correct) or negative (incorrect) across 13 iterations. **Freq.**: proportion of instances with positive/negative contribution.

	Correct		Incorrect		Feature Category					
Q1 Features	**Mean**	**Freq.**	**Mean**	**Freq.**	**Lex.**	**PoS**	**Syn.**	**Aff.**	**Hdg.**	**Unc.**
Punctuation Density	0.170	53.9%	−0.089	46.1%	X					
Function Word Ratio	−0.915	53.9%	0.846	46.1%			X			
Verb Ratio	−0.705	53.6%	0.683	46.4%		X				
Fear	0.580	52.6%	−0.499	47.4%				X		
Noun Ratio	−0.478	52.4%	0.698	47.6%		X				
Sentence Count	−0.510	51.2%	0.531	48.8%	X					
Type-Token Ratio	0.561	49.7%	−0.497	50.3%	X					
Anger	1.196	31.3%	−0.589	68.7%				X		
Total by Category					**3**	**2**	**1**	**2**	**0**	**0**

5.3 *None* Predictions

For *none*, model behavior is shaped by the absence of strong emotional or stylistic markers (Table 6; Figure 3, $\rho = 0.708$). Higher fear was associated with misclassifications (+1.148, 39.2%), whereas lower fear values were strongly linked to correct predictions (−0.520, 60.8%). Unlike *against*, where both anger and fear played a role, fear was the only affective Q1 feature for *none*.

Correct predictions were associated with shorter, more uniform word forms: lower average word length (−0.777, 53.3%) and lower standard deviation of word length (+0.644, 55.0%). In contrast to both *favor* and *against*, lexical diversity played a distinctive role: correct predictions were linked to higher hapax legomena (+0.670, 53.8%) and higher noun ratios (+0.712, 48.1%), suggesting descriptive or factual language. However, the type–token

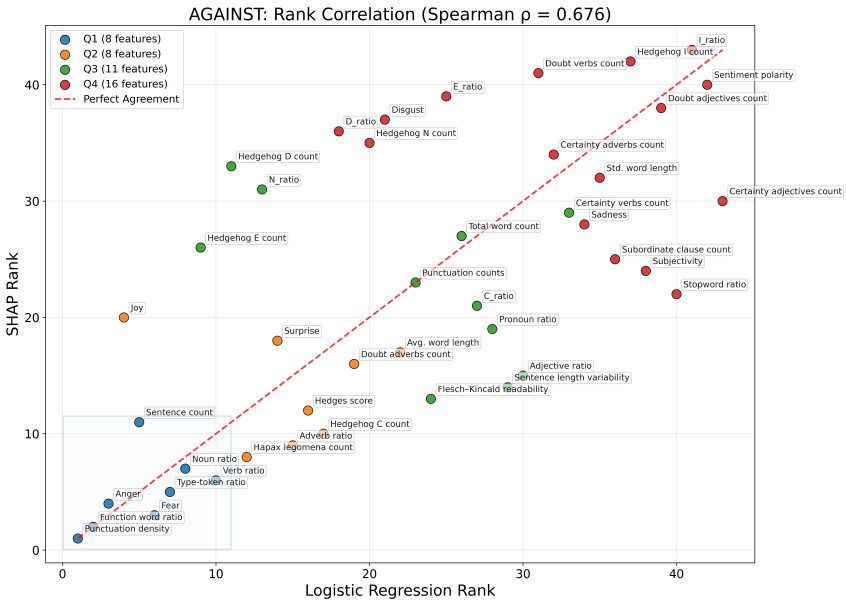

Figure 2: Rank correlation between LR coefficients and SHAP importance for *against* ($\rho = 0.676$).

ratio showed an inverse relationship—lower values were associated with correct predictions (-0.957, 43.0%)—indicating that excessive vocabulary diversity may make it harder for models to identify the absence of stance.

These results suggest that models identify *none* through the absence of stance-related cues. Neutral language—marked by short, consistent word forms, moderate lexical variety, and minimal affect—was linked to accurate predictions. The *none* class is thus defined less by distinctive markers and more by what is missing.

Table 6: Directional influence of Q1 features for *none*. **Mean**: average standardized feature value when SHAP contribution was positive (correct) or negative (incorrect) across 13 iterations. **Freq.**: proportion of instances with positive/negative contribution.

	Correct		Incorrect		Feature Category				
Q1 Features	**Mean**	**Freq.**	**Mean**	**Freq.**	**Lex.**	**PoS**	**Syn.**	**Aff.**	**Hdg.**	**Unc.**
Fear	−0.520	60.8%	1.148	39.2%				X		
Std. Dev. Word Length	0.644	55.0%	−0.854	45.0%	X					
Hapax Legomena	0.670	53.8%	−0.894	46.2%	X					
Average Word Length	−0.777	53.3%	0.840	46.7%	X					
Function Word Ratio	0.193	48.7%	−0.171	51.3%			X			
Noun Ratio	0.712	48.1%	−0.673	51.9%		X				
Type-Token Ratio	−0.957	43.0%	0.628	57.0%	X					
Total by Category					**4**	**1**	**1**	**1**	**0**	**0**

6 Discussion

Figure 4 presents the Q1 features across all stance classes, revealing that each class is shaped by distinctive linguistic patterns. We found only limited overlap between *favor* and *against*: positive mean values for `type-token ratio`, negative mean values for `noun ratio`, and negative mean values for `sentence count` were the only features shared across classes when samples were consistently classified correctly. Collectively, they suggest that conciseness (fewer sentences), reduced reliance on nouns, and greater lexical diversity may serve as general cues for more reliable stance detection, even as the overall profiles of each stance class remain largely distinct. Importantly, these features are not contradictory: shorter texts can still be lexically diverse if they use a wide range of words without repetition, and a lower noun ratio does not mean fewer unique words overall but instead reflects greater reliance on verbs, adjectives, and adverbs that often carry evaluative meaning.

Within *favor*, correctly classified samples were associated with negative values for `hapax legomena`, `punctuation density`, `pronoun ratio`, `word count`, `stopword ratio`, `doubt adverbs`, and `certainty verbs`. Together, these patterns point toward a style of expression that is concise, less repetitive, and less marked by structural or functional fillers. In practice, this suggests that support is often communicated through direct statements that avoid unnecessary elaboration, hedging, or overuse of personal reference.

In *against*, a different set of features characterized correctly classified samples: positive values for `punctuation density`, `fear`, and `anger`, alongside negative values for `function word ratio` and `verb ratio`. This profile suggests that

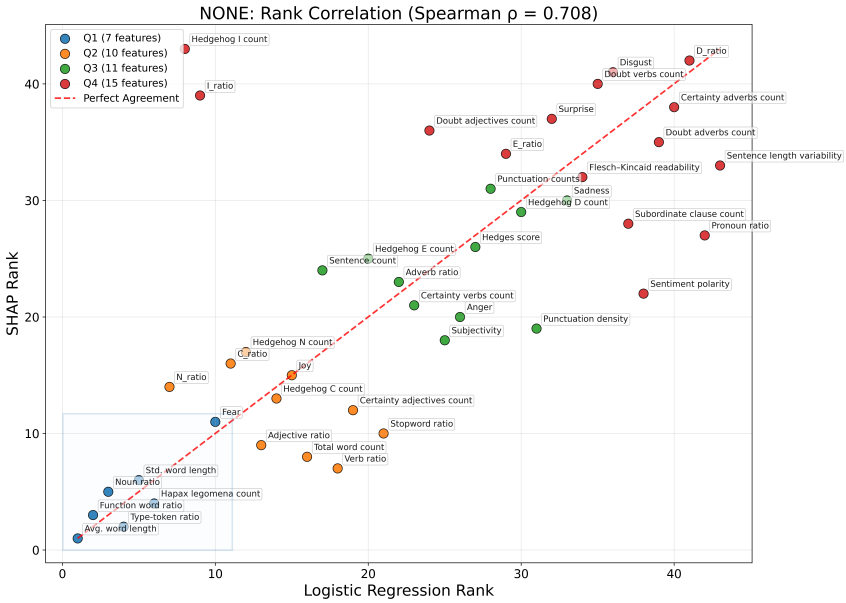

Figure 3: Rank correlation between LR coefficients and SHAP importance for *none* ($\rho = 0.708$).

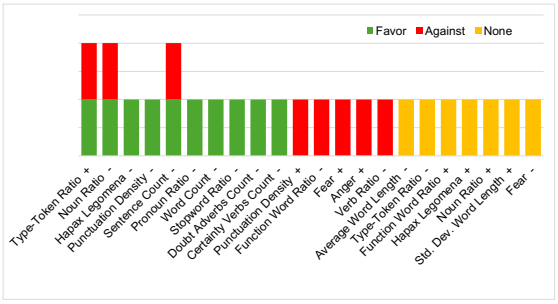

Figure 4: Top-ranked (Q1) features across all stance classes. The sign (+/–) indicates the sign of the mean feature value computed over correctly classified instances only.

opposition is expressed through more marked and emphatic language: greater use of punctuation may signal intensity, while emotional markers reinforce the affective tone of disagreement. A lower reliance on function words and verbs indicates that oppositional stances may be articulated with a focus on content-heavy or affect-laden vocabulary rather than structural connectives. These findings contrast with *favor*, where correct classifications were associated with reduced punctuation and minimal emotional signaling. The two classes highlight complementary stylistic pathways: *favor* is conveyed more reliably through direct and streamlined phrasing, while *against* often depends on heightened emotional expression and structural emphasis.

The *none* class displayed a distinct profile set apart from both *favor* and *against*. Correct classifications were linked to negative values for `average word length`, `type-token ratio`, and `fear`, alongside positive values for `function word ratio`, `hapax legomena`, `noun ratio`, and `std. dev. of word length`. This combination suggests that neutral texts are characterized by shorter and simpler words, greater use of function words, and a somewhat more variable mix of word lengths. The presence of more hapax legomena and a higher noun ratio may indicate descriptive or factual language that does not carry the evaluative signals seen in other stance categories. The absence of emotional markers such as fear appears central to maintaining neutrality; when strong emotions are present, misclassification becomes more likely. In contrast to the streamlined phrasing of *favor* and the affect-laden intensity of *against*, the *none* category is defined less by distinctive stance cues and more by the relative absence of evaluative or emotional markers. Despite this, the *none* class has been largely overlooked in prior stance detection research, and even absent in some datasets (Li et al., 2021), where most baseline studies either exclude it entirely or treat it as a residual category of no analytical interest. Our findings suggest this is a missed opportunity: *none* exhibits its own coherent stylistic profile, and understanding what distinguishes neutral language from subtle stance expressions is essential for reducing false positives across all classes.

Taken together, the three stance classes reveal complementary stylistic pathways that models exploit for classification. Notably, `noun ratio`

and `sentence count` emerged as Q1 features across multiple classes but with opposing directional effects: lower noun ratios aided both *favor* and *against* detection, yet higher noun ratios supported *none* classification, suggesting that descriptive, noun-heavy language reads as factual rather than evaluative. This cross-class reversal underscores that stance detection is not a single classification problem but a family of linguistically distinct sub-tasks, each governed by different stylistic expectations. Systems that treat all stance classes uniformly risk overlooking these asymmetries.

6.1 Related Work

Our findings connect to several lines of prior research. In opinion mining, hand-crafted stylistic features have proven informative (Ahuja et al., 2019; Conroy et al., 2015), and work on fake news has found that deceptive content tends to use simpler, more direct language (Carrasco-Farré, 2022)—paralleling our finding that straightforward texts are more reliably classified. In argument mining, Quensel et al. (2025) showed that hedging exerts complex, context-dependent influences on perceived argument strength.

Within stance detection, Schuff et al. (2017) found clear correspondences between emotion and stance categories, consistent with our results. Prior work has leveraged such patterns in model design (Zhang et al., 2020b; Upadhyaya et al., 2023), but has largely treated affective features as uniformly beneficial. Our findings challenge this by revealing class-dependent effects: fear and anger enhance *against* predictions but degrade accuracy for *none*—asymmetric behavior not systematically documented before.

Hedging cues are well established in related domains (Liu et al., 2024; Quensel et al., 2025) but underexplored in stance detection. In our study, moderate hedging accompanied clear but cautious stances, whereas excessive hedging led to confusion with *none*. In sentiment analysis, hedging typically only softens sentiment strength; in stance detection, hedging can obscure the stance target itself, causing *against* to resemble *none*. This distinction between functional and stance-obscuring hedging has not been explicitly demonstrated in prior work. Similarly, higher lexical diversity often correlated with poorer classification—suggesting that varied vocabulary dilutes stance indicators, paralleling findings in fake news detection (Carrasco-Farré, 2022).

6.2 Limitations and Future Work

This work has several limitations. Our analysis focuses on instance-level features and does not incorporate broader discourse context or higher-level semantics such as world knowledge. Our analysis also treats features individually; interactions among them were not explicitly modeled. Given that stylistic features can be correlated (e.g., word count and sentence count), examining feature interactions and joint effects would be a valuable extension. Findings are also based solely on English data, and class imbalance required undersampling. Disagreement cases across architectures were excluded from our consensus-based analysis. Our multi-target design does not capture per-target or per-dataset variation, so we cannot fully rule out that some stylistic patterns partially reflect dataset- or topic-specific characteristics. Finally, we restricted our experiments to a subset of neural architectures and did not include recent large language models. Future work could address these gaps by examining stance at the dialogue level, integrating semantic knowledge, expanding to non-English datasets, conducting per-target and per-dataset analyses, and applying the framework to larger pretrained and generative models such as LLMs.

7 Conclusion

This study examined how linguistic style influences neural stance detection reliability. Using 43 stylistic features spanning syntactic, lexical, emotional, and epistemic dimensions, we assessed their impact on classification outcomes across six architectures, focusing on consensus samples to ensure stability in observed patterns.

Our results show that each stance class exhibits a distinct stylistic profile: *favor* is more reliably detected when expressed in direct, unhedged language; *against* when marked by strong negative emotion and lexical variety; and *none* when texts are emotionally neutral and lexically simple. Errors were systematically linked to stylistic complexity, emotional ambiguity, or indirect phrasing, indicating reliance on surface-level signals. These findings demonstrate that errors are not random but tied to predictable language patterns, and that future systems will need greater sensitivity to discourse structure, hedging, and pragmatic cues.

Ethics Statement

This study uses four publicly available, previously published stance detection datasets (Mohammad et al., 2016; Li et al., 2021; Glandt et al., 2021; Conforti et al., 2020), all of which were collected and annotated under the ethical guidelines established by their original authors. We did not collect any new data or interact with human subjects. Our analysis focuses on aggregate stylistic patterns across texts and does not attempt to identify, profile, or target individual users. All models were trained and evaluated on existing benchmark data, and our findings are intended to improve the interpretability and robustness of stance detection systems rather than to enable surveillance, censorship, or manipulation of public discourse. We acknowledge that stance detection technology, like other opinion mining tools, carries risks of misuse, and we encourage its application in contexts that respect individual privacy and freedom of expression.

References

Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. 2019. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152:341–348. International Conference on Pervasive Computing Advances and Applications- PerCAA 2019.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

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

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. 2024. Measuring the robustness of NLP models to domain shifts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 126–154, Miami, Florida, USA. Association for Computational Linguistics.

Carlos Carrasco-Farré. 2022. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9(1):1–18.

Meghna Chaudhary and Tempestt Neal. 2026. Implicit aspect extraction: A systematic review. *ACM Comput. Surv.*, 58(7).

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2025. Prompting and fine-tuning open-sourced large language models for stance classification. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence, AUS. IJCAI International Joint Conference on Artificial Intelligence2017, Pages 3988-399426th International Joint Conference on

Artificial Intelligence, IJCAI 2017; Melbourne; Australia; 19 August 2017 through 25 August 2017; Code 130864.

Omar S. Fahmy, Kamel A. Elhaddad, Khaled M. Badran, and Mohamed K. Elhadad. 2025. Enhancing textual deception detection: A fused handcrafted feature approach with machine learning models. In *Machine Learning and Soft Computing*, pages 216–230, Singapore. Springer Nature Singapore.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Comput. Surv.* Just Accepted.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87, Cham. Springer International Publishing.

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

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Jochen Hartmann. 2022. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

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

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2738–2747, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024. Emotion detection for misinformation: A review. *Information Fusion*, 107:102300.

Steven Loria. 2018. Textblob: Simplified text processing. Release 0.15.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Viktor Pekar, Marina Candi, Ahmad Beltagui, Nikolaos Stylos, and Wei Liu. 2024. Explainable text-based features in predictive models of crowdfunding campaigns. *Annals of Operations Research*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1868–1873, Minneapolis, Minnesota. Association for Computational Linguistics.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7).

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A multi-task model for sentiment aided stance detection of climate change tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):854–865.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Hsin-Chang Yang, Yi-Ling Hung, and Ling-Ciao Wang. 2024. Stylometry-based fake news classification using text mining techniques. In *Proceedings of the 2024 11th Multidisciplinary International Social Networks Conference*, MISNC '24, page 85–94, New York, NY, USA. Association for Computing Machinery.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020a. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020b. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Yiwei Zhou, Alexandra I. Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Web Information Systems Engineering – WISE 2017*, pages 18–32, Cham. Springer International Publishing.

8 Appendix

Hyperparameter Search Space

Table 7 details the hyperparameter search space used during Optuna optimization for each model family.

Table 7: Hyperparameter search space for different model types.

Hyperparameter	RNN Models	CNN (KimCNN)	BERT
Epochs	25–200	25–200	3–10
Batch size	16, 32, 64, 128	16, 32, 64, 128	8, 16, 32
Learning rate	10^{-5}–10^{-3} (log)	10^{-5}–10^{-3} (log)	10^{-6}–5×10^{-5} (log)
Dropout	0.1–0.5	0.1–0.5	0.1–0.5
Hidden size	256, 512, 768	—	256, 512, 768
Number of layers	1–3	—	—
FC layer size	—	64, 128, 256	—

9 Full Analysis for *Favor* Stance

A comprehensive directional analysis of all 43 features for the *favor* stance is provided in Table 8.

10 Full Analysis for *Against* Stance

A comprehensive directional analysis of all 43 features for the *against* stance is provided in Table 9.

11 Full Analysis for *None* Stance

A comprehensive directional analysis of all 43 features for the *none* stance is provided in Table 10.

Table 8: Full directional influence analysis of all features for the *favor* stance, grouped by consensus importance quartiles (Q1–Q4).

	Correct Classification		*Incorrect Classification*		**Feature Category**					
Stylistic Feature	**Mean**	**Frequency (%)**	**Mean**	**Frequency (%)**	**Lex. Rich.**	**PoS**	**Synt/Struct**	**Affect**	**Hedging**	**Uncertainty**
Quartile 1 Features										
Doubt Adverbs Count	-0.445	73.2%	0.918	26.8%					**X**	
Certainty Verbs Count	-0.241	66.5%	0.477	33.5%					**X**	
Noun Ratio	-0.656	59.7%	0.647	40.3%		**X**				
Pronoun Ratio	-0.248	55.3%	0.013	44.7%		**X**				
Sentence Count	-0.036	53.3%	0.320	46.7%	**X**					
Punctuation Density	-0.113	52.2%	0.248	47.8%	**X**					
Total Word Count	-0.317	50.7%	0.364	49.3%	**X**					
Hapax Legomena	-0.131	50.5%	-0.006	49.5%	**X**					
Stopword Ratio	-0.645	49.4%	0.518	50.6%	**X**					
Type-Token Ratio	0.010	49.0%	-0.135	51.0%	**X**					
Total by Feature Category					**6**	**2**	**0**	**0**	**2**	**0**
Quartile 2 Features										
Certainty Adverbs Count	-0.235	78.7%	0.789	21.3%					**X**	
Anger	-0.325	74.4%	0.414	25.6%				**X**		
Function Word Ratio	-0.464	52.5%	0.439	47.5%			**X**			
Adjective Ratio	-0.510	51.0%	0.520	49.0%		**X**				
Average Word Length	-0.060	50.5%	0.031	49.5%	**X**					
Hedge C Count	-0.197	49.1%	0.113	50.9%					**X**	
Verb Ratio	0.155	48.5%	0.040	51.5%		**X**				
Fear	1.026	46.5%	-0.431	53.5%				**X**		
Total by Feature Category					**1**	**2**	**1**	**2**	**2**	**0**
Quartile 3 Features										
D Ratio	-0.038	52.6%	0.069	47.4%						**X**
Surprise	0.098	52.0%	-0.116	48.0%				**X**		
Flesch-Kincaid Readability Score	0.107	50.9%	-0.063	49.1%			**X**			
Adverb Ratio	-0.058	49.8%	0.092	50.2%		**X**				
Std. Dev. Word Length	-0.078	47.8%	0.298	52.2%	**X**					
Punctuation Counts	-0.102	47.8%	0.060	52.2%			**X**			
Sentiment Polarity	1.319	33.9%	-0.538	66.1%				**X**		
Hedge E Count	0.159	33.4%	-0.179	66.6%					**X**	
Hedge D Count	0.276	33.2%	-0.063	66.8%					**X**	
Joy	0.896	30.1%	-0.472	69.9%				**X**		
Total by Feature Category					**1**	**1**	**2**	**3**	**2**	**1**
Quartile 4 Features										
Hedge N Count	-0.101	73.3%	0.477	26.7%					**X**	
C Ratio	0.292	69.3%	-0.518	30.7%						**X**
Disgust	-0.086	66.1%	0.136	33.9%				**X**		
N Ratio	-0.043	68.5%	0.291	31.5%						**X**
Doubt Adjectives Count	-0.191	65.4%	0.382	34.6%					**X**	
Sadness	-0.203	61.1%	0.154	38.9%				**X**		
Doubt Verbs Count	0.021	52.9%	0.078	47.1%					**X**	
E Ratio	-0.138	52.8%	-0.074	47.2%						**X**
Hedges Score	-0.032	52.5%	0.044	47.5%					**X**	
Sentiment Subjectivity	-0.179	52.4%	0.232	47.6%				**X**		
Subordinate Clause Ratio	0.251	47.8%	0.108	52.2%			**X**			
Sentence Length Variation	0.121	47.2%	-0.020	52.8%			**X**			
Certainty Adjectives Count	0.376	41.3%	-0.203	58.7%					**X**	
I Ratio	15.070	0.2%	-0.066	46.7%						**X**
Hedge I Count	15.000	0.2%	-0.066	46.7%					**X**	
Total by Feature Category					**0**	**0**	**2**	**3**	**6**	**4**

Table 9: Full directional influence analysis of all features for the *against* stance, grouped by consensus importance quartiles (Q1–Q4).

	Correct Classification		*Incorrect Classification*		**Feature Category**					
Stylistic Feature	**Mean**	**Frequency (%)**	**Mean**	**Frequency (%)**	**Lex. Rich.**	**PoS**	**Synt/Struct**	**Affect**	**Hedging**	**Uncertainty**
Quartile 1 Features										
Punctuation Density	0.170	53.9%	-0.089	46.1%	**X**					
Function Word Ratio	-0.915	53.9%	0.846	46.1%			**X**			
Verb Ratio	-0.705	53.6%	0.683	46.4%		**X**				
Fear	0.580	52.6%	-0.499	47.4%				**X**		
Noun Ratio	-0.478	52.4%	0.698	47.6%		**X**				
Sentence Count	-0.510	51.2%	0.531	48.8%	**X**					
Type-Token Ratio	0.561	49.7%	-0.497	50.3%	**X**					
Anger	1.196	31.3%	-0.589	68.7%				**X**		
Total by Feature Category					**3**	**2**	**1**	**2**	**0**	**0**
Quartile 2 Features										
Joy	-0.150	59.3%	-0.016	40.7%				**X**		
Adverb Ratio	-0.437	57.6%	0.343	42.4%		**X**				
Surprise	-0.131	55.8%	0.210	44.2%				**X**		
Doubt Adverbs Count	-0.182	55.3%	-0.105	44.7%					**X**	
Hapax Legomena	-0.702	50.6%	0.678	49.4%	**X**					
Average Word Length	0.215	48.8%	0.105	51.2%	**X**					
Hedge C Count	0.285	45.6%	-0.152	54.4%					**X**	
Hedges Score	0.447	29.2%	-0.349	70.8%					**X**	
Total by Feature Category					**2**	**1**	**0**	**2**	**3**	**0**
Quartile 3 Features										
C Ratio	0.413	80.9%	-1.170	19.1%						**X**
N Ratio	-0.225	80.3%	0.776	19.7%						**X**
Hedge D Count	-0.228	76.0%	0.163	24.0%					**X**	
Certainty Verbs Count	-0.387	65.8%	0.329	34.2%					**X**	
Adjective Ratio	-0.629	56.9%	0.833	43.1%		**X**				
Sentence Length Variation	-0.047	52.6%	0.132	47.4%			**X**			
Flesch-Kincaid Readability Score	0.564	51.8%	-0.293	48.2%			**X**			
Pronoun Ratio	-0.269	50.2%	0.053	49.8%		**X**				
Total Word Count	0.058	49.0%	-0.029	50.7%	**X**					
Punctuation Counts	-0.044	47.9%	-0.043	52.1%			**X**			
Hedge E Count	0.478	26.6%	-0.196	73.4%					**X**	
Total by Feature Category					**1**	**2**	**3**	**0**	**3**	**2**
Quartile 4 Features										
I Ratio	-0.074	91.5%	0.627	8.5%						**X**
E Ratio	-0.216	79.9%	0.615	20.1%						**X**
Hedge I Count	-0.062	77.0%	0.192	23.0%					**X**	
Doubt Verbs Count	-0.158	73.6%	0.080	26.4%					**X**	
Hedge N Count	-0.079	58.7%	0.101	41.3%					**X**	
Doubt Adjectives Count	-0.098	54.7%	0.006	45.3%					**X**	
Sentiment Polarity	-0.160	54.4%	0.141	45.6%				**X**		
D Ratio	-0.108	52.9%	-0.151	47.1%						**X**
Sentiment Subjectivity	-0.160	50.5%	0.038	49.5%				**X**		
Std. Dev. Word Length	0.141	50.5%	0.127	49.5%	**X**					
Stopword Ratio	0.076	50.3%	-0.126	49.7%	**X**					
Subordinate Clause Ratio	0.499	47.9%	-0.462	52.1%			**X**			
Disgust	-0.007	46.8%	0.049	53.2%				**X**		
Certainty Adjectives Count	0.624	37.4%	-0.369	62.6%					**X**	
Certainty Adverbs Count	0.247	33.1%	-0.199	66.9%					**X**	
Sadness	1.128	29.4%	-0.367	70.6%				**X**		
Total by Feature Category					**2**	**0**	**1**	**4**	**6**	**3**

Table 10: Full directional influence analysis of all features for the *none* stance, grouped by consensus importance quartiles (Q1–Q4).

	Correct Classification		*Incorrect Classification*		**Feature Category**					
Stylistic Feature	**Mean**	**Frequency (%)**	**Mean**	**Frequency (%)**	**Lex. Rich.**	**PoS**	**Synt/Struct**	**Affect**	**Hedging**	**Uncertainty**
Quartile 1 Features										
Fear	-0.520	60.8%	1.148	39.2%				**X**		
Std. Dev. Word Length	0.644	55.0%	-0.854	45.0%	**X**					
Hapax Legomena	0.670	53.8%	-0.894	46.2%	**X**					
Average Word Length	-0.777	53.3%	0.840	46.7%	**X**					
Function Word Ratio	0.193	48.7%	-0.171	51.3%			**X**			
Noun Ratio	0.712	48.1%	-0.673	51.9%		**X**				
Type-Token Ratio	-0.957	43.0%	0.628	57.0%	**X**					
Total by Feature Category					**4**	**1**	**1**	**1**	**0**	**0**
Quartile 2 Features										
Hedge N Count	-0.197	95.6%	4.589	4.4%					**X**	
Stopword Ratio	0.562	51.7%	-0.601	48.3%	**X**					
Verb Ratio	-0.792	50.9%	0.825	49.1%		**X**				
Hedge C Count	0.240	50.3%	-0.103	49.7%						
Total Word Count	-0.065	49.6%	0.314	50.4%	**X**					
Adjective Ratio	0.847	43.8%	-0.653	56.2%		**X**				
Joy	1.318	23.9%	-0.477	76.1%				**X**		
Certainty Adjectives Count	2.137	19.4%	-0.437	80.6%					**X**	
C Ratio	-1.948	14.4%	0.388	85.6%						**X**
N Ratio	4.414	4.4%	-0.193	95.6%						**X**
Total by Feature Category					**2**	**2**	**0**	**1**	**2**	**2**
Quartile 3 Features										
Hedge D Count	-0.180	91.5%	1.699	8.5%					**X**	
Anger	-0.396	76.1%	1.299	23.9%				**X**		
Hedge E Count	-0.179	65.6%	0.210	34.4%					**X**	
Sentence Count	-0.320	58.5%	0.668	41.5%	**X**					
Sentiment Subjectivity	-0.092	52.2%	0.088	47.8%				**X**		
Certainty Verbs Count	0.142	51.4%	-0.007	48.6%					**X**	
Adverb Ratio	0.175	50.6%	-0.167	49.4%		**X**				
Punctuation Density	0.254	48.7%	-0.323	51.3%	**X**					
Punctuation Count	0.268	45.9%	-0.286	54.1%			**X**			
Hedges Score	0.362	35.6%	-0.169	64.4%					**X**	
Sadness	0.991	28.3%	-0.353	71.7%				**X**		
Total by Feature Category					**2**	**1**	**1**	**3**	**4**	**0**
Quartile 4 Features										
Certainty Adverbs Count	-0.247	79.3%	0.698	20.7%					**X**	
Hedge I Count	-0.060	78.3%	0.136	16.1%					**X**	
Doubt Adjectives Count	-0.268	78.3%	0.830	21.7%						
E Ratio	-0.186	67.7%	0.210	32.3%						**X**
Subordinate Clause Ratio	-0.406	56.1%	0.494	43.9%			**X**			
D Ratio	-0.043	52.6%	-0.017	47.4%						**X**
Sentence Length Variation	-0.142	52.4%	0.267	47.6%			**X**			
Flesch-Kincaid Readability Score	0.167	48.3%	-0.198	51.7%			**X**			
Surprise	0.126	44.4%	-0.248	55.6%				**X**		
Sentiment Polarity	0.361	40.5%	-0.225	59.5%				**X**		
Pronoun Ratio	0.539	40.3%	-0.326	59.7%		**X**				
Doubt Adverbs Count	0.694	35.9%	-0.279	64.1%						
Disgust	0.153	35.8%	-0.005	64.2%				**X**		
I Ratio	0.138	31.8%	-0.073	60.4%						**X**
Doubt Verbs Count	0.223	24.7%	-0.142	75.3%					**X**	
Total by Feature Category					**0**	**1**	**3**	**3**	**3**	**3**