# Explainable Subjective Stance Classification with SetFit in Political Discourse

**Institute of Computer Science, Brandenburgische Technische Universität Cottbus-Senftenberg**

Juan-Francisco Reyes

pacoreyes@protonmail.com

Stance classification in NLP is not just an academic exercise, but a crucial tool for understanding political discourse and the attitudes underlying political statements. This research addresses the challenge of limited annotated datasets in political science by proposing a practical sentence-level dataset for binary subjective stance classification—support or oppose—using the SetFit few-shot learning framework. The study leverages the Sentence Transformers architecture and incorporates traditional linguistic approaches to enhance explainability. We employ corpus linguistics, tailored lexicons, and lexicogrammatical rules to identify key linguistic features such as pro/con polarity, affective and epistemic dimensions, and modality markers. SHAP analysis quantifies the influence of these features on SetFit model decisions. Our findings demonstrate the efficacy of few-shot learning in subjective stance classification and highlight the importance of linguistic features, particularly pro/con polarity and affective expressions. The StanceSentences dataset and our hybrid analytical approach offer a benchmark for future research, emphasizing the need for nuanced, multi-layered analysis in political discourse.

*Keywords:* stance classification, political discourse analysis, SetFit, explainable AI, SHAP.

## 1. Introduction

*Stance classification* in *Natural Language Processing* (*NLP*) plays a pivotal role in deciphering the intricacies of political discourse and uncovering underlying intentions and attitudes toward various topics. This research addresses significant challenges in this subject, including the limited availability of annotated datasets for stance classification, a notable concern in specialized areas like political science. This limitation arises from the inherent complexity in manually compiling political statements that express *stance*, which can be implicit, subtle, indirect, or even hidden. Moreover, the existing datasets for stance detection or classification in English-language primarily compile multi-sentence texts, specifically *tweets* from online political conversations, creating a shortfall in research on NLP models designed for sentence-level stance classification within *political discourse analysis* (*PDA*), and particularly in analyzing the language used by actual politicians. We focus on sentences that express *subjective stance* from direct to moderately subtle political language. This pragmatic approach permits us to create a foundational understanding of the core linguistics of stance and evaluate the methodology's effectiveness and the clarity of the resulting artifacts. We utilized sentences about *political issues* that have been in the American public sphere in recent years, providing a broad spectrum of linguistic patterns to represent ideological viewpoints.

In response to these challenges, we propose the construction of a sentence-level dataset for binary stance classification—*support* (*pro*/*favor*) or *oppose* (*con*/*against*)—using a *bootstrapping* method within the *few-shot learning* framework of *SetFit* (Tunstall et al., 2022) suitable when labeled data is limited, to iteratively refine the model. SetFit's methodology aligns perfectly with the challenges of collecting and annotating sentences of political discourse texts that contain stance expressions, determining whether it can efficiently analyze and understand complex political language with limited data. SetFit, with its advanced use of the *Sentence Transformers* architecture (Reimers & Gurevych, 2019), should be proficient at overcoming these subtleties, decoding everything from explicit statements to the more nuanced shades of expression, ensuring a thorough and comprehensive analysis.

Because of the concerns about the explainability of neural models' predictions, this study aims to go back to more traditional linguistic approaches—corpus linguistics methods, tailored lexicons, and lexicogrammatical rules—to decode how neural models prioritize and utilize specific word choices for stance classification. By focusing on adjectives, adverbs, and verbs categorized into eight distinct linguistic features—*positive affect*, *negative affect*, *pro polarity*, *con polarity*, *certainty*, *emphatics*, *doubt*, and *hedges*— we mitigates bias excluding nouns and *named entities* to understand stance classification without the liability of thematic content. Therefore, we bridge the empirical performance of neural models with the nuanced understanding and interpretability that linguistic analysis provides at different levels of explainability: (1) a transparent *linguistic-rule-based model*, *(LRBM)* operationalized on different levels of linguistic structures via *spaCy* (Honnibal & Montani, 2017), and (2) *a SHAP* (*SHapley Additive exPlanations*) (Nohara et al., 2019) analysis to dissect the impact of these linguistic features on the SetFit model's decision-making process, enabling us to quantify the influence of stance features on stance classification outcomes.

This study aims to answer key questions about stance classification in the domain of political discourse:

- *RQ1*: How does the SetFit model performance behave along each iteration of the bootstrapping process while building a stratified dataset for stance classification at the sentence level in political discourse?
- *RQ2*: What linguistic features are most important to predict the classification of support or oppose stance?

The contribution of this work is three-fold: (1) the introduction of a dataset for few-shoot learning models for classifying binary stance expressions in political discourse; (2) a quantitative analysis of the SetFit model performance along the bootstrapping process by tracking metrics; and (3) an explainability qualitative and quantitative framework that combines a transparent feature-based model with SHAP analysis of the SetFit model behavior for a comprehensive analysis of the linguistic features of stance classification.

# 2. Related Work

(This literature review will indistinctly peruse stance classification and stance detection as they are closely related.) Stance in linguistics broadly refers to the expression of a speaker's attitude, feelings, evaluations, or commitment towards a proposition or an entity (Biber et al., 1999), encompassing a range of linguistic mechanisms through which speakers position

themselves relative to their utterances and their interlocutors. The *stance triangle*, proposed by Du Bois (2007), is a fundamental concept in stance analysis that proposes stance as a relational act composed of three dynamically assembled components:

- *The speaker* (or *stance-taker*): The individual who expresses stance.
- *The target* (or *stance object*): The person, an idea, a situation, or any other entity about which the stance is addressed.
- The *addressee* (or *stance audience*): The individual or group to whom the stance is communicated and whose reactions can significantly influence how the stance is expressed.

The stance object, or "*target*", can be represented in two forms: (1) *noun-phrase*, a more straightforward representation where the target is a specific entity or a set of entities described by a noun phrase, for instance, "*the new tax policy is unfair*" the noun phrase target is "*the new tax policy*"; or (2) *claim*, a broader statement, opinion, or assertion that can be agreed or disagreed with, for example, in the statement "*Implementing AI in car driving will reduce accidents*." the target is the claim itself. Similarly, stance may be addressed to (1) *multitarget* when acknowledges opinions towards different entities or as (2) *target-specific* when the focus is towards a single target (Du et al., 2017; Sobhani et al., 2017).

In the domain of political communication, stance assumes a nuanced role, owing to the complexity and strategic nature of political discourse. The study of political stance is deeply rooted in the understanding that political language is not just a medium of communication, but a potent tool of persuasion and ideological expression (Chilton, 2004). It is through this language that politicians shape public opinion, assert power, and negotiate identities (Wilson, 1990). Martin and White (2005) delved into the role of stance in political communication, elaborating on the *appraisal theory*, which provides profound insights into how language is harnessed to evaluate issues and, consequently, take a stance in political texts. Their framework has become indispensable for dissecting the ways in which politicians express attitudes, make judgments, and interact with audiences.

In *Styles of stance in English: Lexical and grammatical marking of evidentiality and affect* (1989), Biber and Finegan explored how lexical and grammatical elements can convey a speaker or writer's attitudes, evaluations, feelings, and perceptions of truth that express stance. Their taxonomy of stance features aids in understanding the multifaceted nature of stance and provides a comprehensive approach to analyzing it in text.

**Table 1**

*Categories of stance features by Biber and Finegan (1989).*

| Category | Description |
|---|---|
| *Affect markers* | Adverbs, verbs, and adjectives that express emotions, evaluations, or attitudes towards the proposition. |
| *Certainty and doubt markers* | Adverbs, verbs, and adjectives that either express epistemic certainty or doubt. |
| *Hedges and emphatics* | Linguistic devices that either downplay or amplify the force of the statement, reflecting the speaker's or writer's commitment |

| Category | Description |
|---|---|
| | to the proposition. |
| *Modal verbs* | Verbs that indicate necessity, possibility, permission, or ability, providing insights into the speaker's perspective on the likelihood or necessity of the proposition. |

Biber and Finegan employed extensive corpus-based methodologies to analyze linguistic features systematically across large datasets. Their methodology used statistical and computational techniques to identify patterns of language use, following an empirical analysis of the frequency, distribution, and co-occurrence patterns of various markers of stance across different texts and genres, and it has been influential in various research fields, including discourse analysis, sociolinguistics, and computational linguistics.

Sentiment valence has been used as a stance marker along with other features to predict stance. Aldayel & Magdy, (2021), Chauhan et al. (2019), Lai et al. (2020a), Mohammad et al. (2016), Sobhani et al. (2016), Somasundaran & Wiebe (2010), Sun et al. (2018, 2019) among others, confirmed that sentiment may be useful for stance detection when combined with other features. However, integrating sentiment valence with other features enhances the accuracy of stance detection, but sentiment alone is insufficient to fully capture stance nuances.

Among the scarce studies on stance detection and classification relying to some extent on *traditional linguistic features*, after the rise of computational linguistics and NLP, we found that Somasundaran and Wiebe (2010) explored arguing-based features such as modal verbs and sentiment valence to predict stance classification. Likewise, Anand et al. (2011) utilized multiple linguistic features and structures to predict stance, like counts, repeated punctuation, and other lexicon-based and dependency-based features. Likewise, Hasan and Ng (2013) developed a method for understanding the stance at the semantics level expressed in sentences from American political discourse, using patterns that analyze both the structure and meaning of language, allowing them to detect the underlying attitudes in a politician's statements. Their technique focuses on how sentences are constructed (*syntactic dependencies*) and the broader contexts they fit into (semantic frames). This approach helps identify stances even when different words are used to express similar opinions. Khamkhien (2014) studied the use of linguistic features in expressing evaluative stance in academic discourse, specifically in research article discussions within applied linguistics and language teaching. Likewise, Sun et al. (2016) explored four linguistic features, including lexical, morphology, semantic, and syntax features in Chinese micro-blogs for stance classification.

More recently, a second group of studies used traditional linguistic features in combination with *statistical or Machine Learning* (*ML*) *features*, which are more abstract representations of language, like *N-grams*, *word-embeddings*, *Term Frequency-Inverse Document Frequency* (*TF-IDF*), *bag of words*, Etc. For instance, Walker et al. (2012) used a combination of linguistic features with abstract features to study stance classification, including affective, rebuttal, unigrams, and topic specific features. Also, Lai et al. (2020a) examined the adaptability of stance detection tools across languages using a multi-lingual model, MultiTACOS, highlighting the significance of considering various linguistic cues beyond mere word sequences, such as sentiment and argumentation. Wang et al. (2020) introduced a hierarchical network that employs an attention mechanism to prioritize different linguistic

inputs and establish mutual attention between documents and their linguistic characteristics. Similarly, Vychegzhanin and Kotelnikov (2021) developed an *Ensemble-based Stance Detection* (*ESD*) strategy to identify an author's viewpoint, focusing on the optimal selection of features, including word and character n-grams, dependency structures, and other relevant linguistic and stylistic indicators.

When the advent of computational linguistics and NLP transformed political discourse analysis, researchers leveraged machine learning and text analytics to understand intricacies and patterns in political language. This computational advance and the availability of specialized datasets enabled the analysis of stance in political communication at a scale and depth previously unattainable. As Alturayeif et al. (2023) point out, the interest in studying stance grew especially from the publication of the *SemEval-2016* (*Semantic Evaluation 2016*) competition, which presented the first benchmarked dataset for stance detection (Mohammad et al., 2016). The shift towards neural models, especially with the emergence of *large language models* (*LLM*), has achieved high empirical performance in different NLP tasks, but at the expense of understanding the linguistic principles behind the model's success. The explainability of linguistic features in NLP models is crucial to understanding how linguistic features impact decision-making (Jurafsky & Martin, 2023). Moreover, the explainability of linguistic features facilitates model debugging, identifies and mitigates biases, and ensures that the system adheres to ethical and legal standards if necessary. Recent NLP research emphasizes the synergy between linguistic features and neural models that enhance transparency and linguistics explainability.

## 2.1 Stance Explanation

The exploration of NLP explainability methods in stance detection and classification has seen significant developments, with various scholars employing different techniques, datasets, and levels of explanation granularity, as seen in Table 2.

**Table 2**

*NLP Studies Using Explainability Methods for Stance Detection and Classification with English Datasets Published After SemEval-2016.*

| Author(s) | Method | Dataset(s) | Explanation granularity |
|---|---|---|---|
| Du et al. (2017) | Target-specific attention mechanism | Semeval-2016 | Token |
| Mohtarami et al. (2018) | Semantic relations | FakeNewsChallenge | Phrase |
| Li and Caragea (2019) | Target-specific attention mechanism | SemEval-2016 | Token |
| Popat et al. (2019) | Incremental contribution analysis | Perspectrum | Phrase |
| Jayaram and Allaway (2021) | Mean attention weights (MAW) | VAST | Passage |
| Kawintiranon and Singh (2021) | Weighted log-odds-ratio | Ad-hoc (1) | Passage |
| Conforti et al. (2022) | Pre-annotated weight | WT–WT | Sentence |
| Zhang et al. (2022) | ChatGPT | SemEval-2016 / P-Stance | Passage |
| Saha et al. (2024) | Argument-relevance weight | Createdebate / Room For Debate | Passage |

*Note:* (1) Dataset created ad-hoc for the research.

Du et al. (2017) and Li and Caragea (2019) used attention weights to explain the contribution of individual tokens to the stance prediction, identifying tokens with the highest weight as important for their explanation. Similarly, Popat et al. (2019) and Mohtarami et al. (2018) focused on phrase-level explanations but diverged in their methods. Popat et al. adopted an incremental contribution analysis approach using the Perspectrum dataset (Chen et al., 2019), a method that quantifies the contribution of individual phrases to the overall stance. Mohtarami et al. explored semantic relations using the FakeNewsChallenge dataset (Hanselowski et al., 2018), investigating how the relationships between different phrases influence in stance detection. Conforti et al. (2022) introduced a pre-annotated weight method, utilizing the WT-WT dataset (Conforti et al., 2020) and providing explanations at the sentence level. This approach differs from phrase-level explanations by examining the weight or significance of entire sentences in determining stance, offering a broader context for interpretation. Further refining in the granularity of explanation, Jayaram and Allaway (2021) applied a pre-annotated weight method to the VAST dataset (Allaway and McKeown, 2020), focusing on the token level. This method drills down to the individual word or token, offering precise insights into how specific words contribute to the detected stance.

From a traditional linguistics perspective, the abovementioned research on stance detection and classification using explainability tools allows linguists to identify specific linguistic features and patterns that models use to determine stance. Syntactic structures, word choices, or semantic relations may provide empirical evidence of how language conveys attitudes and beliefs and relates knowledge to linguistic theories related to pragmatics, discourse analysis, and sociolinguistics. Also, AI explainability tools can help uncover biases in stance detection and classification models, especially in the context of automated NLP systems that need close follow-up of their decisions.

## 2.2 Datasets for Stance Detection and Classification

Reviewing datasets built for stance detection and classification in the domain of politics in English language (Table 3), we observed patterns in how stance has been studied:

1. *Preference for tweets*: The preference for tweets as the main material for stance datasets may reflect the easy availability and volume of data in the Twitter API, which offers (1) easy availability of structured data, including metadata with timestamps, engagement metrics, Etc; (2) brevity and focus of text, making them suitable for studying rhetorics or argumentation in political conversations and debates in the public sphere—where stance is prevalent; and (3) readiness of use, since the data needs minimal effort in preprocessing and cleaning.
2. *Passage-level focus*: Since all datasets share the same source—the *X API* (formerly *Twitter API*)—the granularity of the examples is at the passage level (multiple sentences).
3. *Large sizes*: The smallest dataset analyzed had around 3,500 examples (Darwish et al., 2017), and the largest 21K examples (Li et al., 2021), making their construction a formidable task. This extensive demand of examples is aligned to the need of "conventional" LLMs, such as BERT or GPT, in opposition to few-shot learning models, such as SetFit, highlighting the complexity and depth of the research process.

4. *Event specificity*: In some cases, datasets for stance are build based on significant political events and eras, such as the 2016 U.S. Presidential Election in "*Trump vs. Hillary*" by Darwish et al. (2017) and the Brexit referendum in "*TW-BREXIT*" by Lai et al. (2020b).
5. *Public opinion focus*: The datasets for stance detection and classification, while comprehensive in their collection of public opinion in the "*Twittersphere*", notably do not capture the voice of professional politicians in the public sphere. This emphasis on public opinion underscores the importance of understanding and analyzing the sentiments of the general public in political discourse.

**Table 3**

*Datasets in English for Stance Detection and Classification in the Political Discourse Domain.*

| Dataset | Author(s) | Granularity* | Size |
|---|---|---|---|
| Semeval-2016 | Mohammad et al. (2016) | Passage | 4,870 tweets |
| Trump vs. Hillary | Darwish et al. (2017) | Passage | 3,450 tweets |
| Multi-target SD | Sobhani et al. (2017) | Passage | 4,455 tweets |
| TW-BREXIT | Lai et al. (2020b) | Passage | 5,400 tweets |
| P-Stance | Li et al. (2021) | Passage | 21,574 tweets |

*Note:* (*) The granularity of examples.

Although single sentence stance datasets and multi-sentence stance examples each offer distinct benefits for research and applications in NLP, the the analysis of stance at the sentence level offers key benefits: (1) sentence-level stance usually consist of clear, concise statements, simplifying the task of annotating and interpreting stance; (2) due to their brevity, single sentence examples can be annotated more quickly, making more feasible creating larger datasets; and (3) sentence-level stance allow researchers to focus more specifically on the linguistic features that convey stance within a standalone statement. This can help in identifying key indicators of stance such as specific word choices, morphosyntactic structures, or rhetorical devices without the complexity introduced by longer contexts. Thus, there is room to experiment with NLP models and datasets at the sentence-level.

## 2.3 SetFit

Introduced by Tunstall et al. (2022) through collaboration between *Intel Labs*, *UKP Lab* (*Ubiquitous Knowledge Processing Lab*), and *Hugging Face*, SetFit is a framework designed to fine-tune pre-trained *Sentence Transformers* models like *BERT*, *RoBERTa*, *or DistilBERT* for specific text classification tasks with limited labeled data. Notably, at 1,600 times smaller than other *LLM*s (*Large Language Models*) like *OpenAI GPT-3*, SetFit enhances performance and scalability without sacrificing performance (Wasserblat, 2021). This efficiency makes SetFit a cost-effective solution for real-world scenarios with scarce data and limited computational power, making it suitable for NLP projects focused on sentence classification.

However, SetFit's benefits also imply a trade-off regarding the depth of linguistic understanding and contextual nuance that larger models, with their extensive training on diverse and voluminous datasets, can offer. The performance of SetFit is heavily reliant not only on the quality of the pre-trained Sentence Transformers it fine-tunes but also on the quality of the dataset used. If these base models are not adequately trained, are biased, or if the dataset quality is poor, SetFit's output will inherit these limitations, underscoring the importance of high-quality datasets and base models. Sentence Transformers is susceptible to nuanced meanings by converting the entire sentence's context into numerical *embeddings* representing semantic content.

Overall, we found several gaps in the literature review: (1) inexistent research on stance classification using few-shot learning models—in general—and SetFit—in particular; (2) existent datasets for stance classification in English are entirely built from online debates on Twitter, and not from political discourse by professional politicians (3) since existent datasets for stance classification in English are built from tweets, the stance analysis extends to the multi-sentence (passage) level and not with the granularity of single sentences; and (4) inexistent research on explainable stance classification using traditional linguistic features and not purely statistical or ML features.

# 3. Models

To simplify the inherent complexity of stance in political discourse and to provide a controlled linguistic environment, facilitating the assessment of model performance and data efficiency, our research specifically focuses on stance expressions that articulate subjective stance statements, both individual and collective, ranging from clear and direct to moderately subtle or implicit, encapsulated in single sentences. Subjective stance expressions are those that employ first-person (for instance, "*I*") or plural first-person (for instance, "*we*") pronouns, expressing views or opinions that reflect the personal alignment or opposition of individuals or groups towards certain ideas. The study of subjective stance offers several distinct advantages, such as (1) access to a more focused analysis of personal and collective viewpoints—central to understanding political discourse; (2) better control for variability and noise in the data, which often arise from more general or ambiguous statements; and (3) assurance that the models we develop were more easy to interpret. For the sake of simplicity, we will refer mostly to subjective stance as "stance," using the complete name of stance when it is strictly necessary throughout the remainder of this article.

In this study, we introduce two datasets: (1) *StanceSentences*, a collection of 1,280 sentences, and (2) *StanceSentencesFeat*, stance features extracted from *StanceSentences;* and three models, (1) *Linguistic-rule-based Model* (*LRBM*), a rule-based model to extract features from text, (2) *StanceFeat*, a Logistic Regression (LR) model fit with the *StanceSentencesFeat* dataset, and (3) *StanceFit*, a SetFit model fine-tuned with the *StanceSentences* dataset.

## 3.1 Datasets

To ensure a robust evaluation framework, we designed StanceSentences to be perfectly balanced, with equal representation of both stance classes, aiming to mitigate bias and increase the generalizability of our findings. To create StanceSentences, we collected public discourses using an ad-hoc web-scraping tool from American targeted websites,

predominantly from *The American Presidency Project* (Peters & Woolley, n.d.) but also news websites, government archives, and government agencies websites. From the collected 97K speeches, interviews, debates, or similar, we filtered sentences that complied with strict characteristics implemented in an automatic filtering system that analyzed each sentence. Overall, the filtering system followed these criteria to select candidate sentences:

1. The match of at least one political issue of 158 political issues that have been prominent in political discussions and the public sphere in the U.S. over the past 80 years. For each issue, the rule-based system used variations of their written form or synonyms, totaling a dictionary of 369 different expressions. The filtering system built on spaCy and rooted on a custom Named-Entity Recognition (NER) component was designed to identify specific terms using three key matchers: (1) *Hyphenated term* pattern, which identifies compound words in its lemma and non-hyphenated forms (for example, "*same-sex marriages*" to its lemmatized version "*same sex marriage*"); (2) *Lemmatized* pattern, which allows the system to recognize different forms of a word as the same entity (for example, "*taxes*" and "*tax*"); and (3) *Exact-term* matching, ensuring precise identification of specific phrases (for example, "*NATO*" and "*N.A.T.O.*").
2. The matched political issue played a significant role in the main topic of the sentence. Its grammatical role was a *subject*, *direct object*, *object of a preposition*, *attribute*, or *adverbial clause modifier*, and its closeness to the sentence's main verb not less than seven tokens away. This evaluation was made using spaCy's linguistic capabilities.
3. The presence of at least one semantic frame, parsed by an instance of *frameBERT* (Li et al., 2023), a BERT-based frame-semantic parser in terms of FrameNet (Ruppenhofer et al., 2016), ensuring the sentence has content with structured meaning.
4. The sentence length with at least five tokens and at most 50 tokens, measured with spaCy's linguistic capabilities.
5. The number of clauses higher than zero but not more than three, measured with spaCy's linguistic capabilities.
6. The absence of other basic quality features for our specific task, such as question forms, leading patterns (speaker/interview/interviewer labels), incompleteness, repeated words, Unicode, Etc.
7. Presence of singular or plural personal pronouns, "*I*", "*me*", "*my*", "*mine*", "*myself*", "*we*", "*us*", "*our*", "*ours*", "*ourselves*", detected with spaCy's linguistic capabilities.

The filtering task resulted in a pool of 14,101 unlabelled sentences analyzed through a bootstrapping approach of ten rounds of annotation and model inferences classifying sentences into support or oppose. Initially (in iteration 1/10), we built our ground truth, a seed dataset with 180 sentences (90 supporting and 90 opposing) and a test dataset with 200 sentences (100 in each class) that we used consistently to test all models fine-tuned through the bootstrapping process. In each iteration, the pool of sentences was submitted to StanceFit for inference, which assigned each sentence a prediction confidence score that we used to create a ranking of sentences, from which a batch of 100 sentences with perfect class balance with the highest score was retained and added to StanceSentences dataset after a subsequent human evaluation phase performed by four annotators. All sentences were ultimately selected by humans, following these strict criteria:

1. Focus on sentences that convey the speaker's (subjective) stance (like, "*I believe that*"), not someone else's stance (like, "*The president believes that...*").
2. The stance target had to be a political issue and not personal pronouns or determiners ("*this*", "*that*", "*those*", "*these*", "*he*", "*she*", "*his*", "*hers*", "*their*", Etc.)
3. Preference for explicit stance expressions ("*we foster*", "*I support*", or "*we have a commitment with*"), but also accepting a certain degree of subtlety using adjectives like "[something/someone] *is critical to*" or "*we ought to be a little tougher on*".
4. Avoidance of ambiguous, obscure, or cryptic stance expressions, like "*I don't think we fully appreciated the degree of corruption that was in the officer ranks in the military.*" or "*At the end of the day, I think Russia is going to be a very big issue, but not the way we think.*"

This process expanded StanceSentences every round and ensured the inclusion of high-quality, verified entries (Figure 1). In order to minimize bias, an anonymous review was conducted by two additional annotators unaware of the initial classifications, with a third curator resolving any disagreements. The *inter-annotator agreement* (*IAA*) achieved a *Cohen's Kappa* score of 0.889. After each round,StanceFit was fine-tuned with the newly augmented dataset and tested against the balanced test dataset, ensuring continual improvement in model performance.

**Figure 1**

*Examples of Subjective Stance Sentences in the StanceSentences Dataset.*

| Stance | Sentence |
|---|---|
| Support | And I firmly believe that excellence in education is going to be the leading edge of change for New Orleans. |
| Support | In addition to bolstering Ukraine's resistance on the battlefield, we are also demonstrating our support for the people of Ukraine. |
| Oppose | But I think that since there is no inflation in the economy, interest rates should not continue to go up. |
| Oppose | At the economic summit meeting at Versailles, we reaffirmed our commitments to the fight against inflation, to expanded trade, and to economic development. |

*Note:* In blue, stance speakers in first-person, and in red, clearly defined stance targets.

The dataset ended up comprising sentences predominantly from American presidents and vice presidents, but also from a small fraction of stance expressions from other political figures and government officials in the international setting from the period 1939 to 2023 and scraped from *The American Presidency Project* (986), *CNN* (174), *Rev.com* (66), *United States Senate* (31), *United States House of Representatives* (10), *ABC News* (6), *USEmbassy.gov* (2), *The New York Times* (1), *National Archives and Records Administration* (1), *UN.org* (1), *The White House* (1), *USA Today* (1), and *The Pueblo Chieftain* (1).

Once StanceSentences was completed, we extracted numerical representations of its linguistic features using the LRBM to create the StanceSentencesFeat dataset. We used the work of Biber and Finegan (1989) as a foundational theoretical framework to build the LRBM, using its detailed description of stance features and lexicons extending or simplifying whenever it was necessary. (We return to this issue in the experimental set-up description and, with more details, in the Appendix).

As seen in Table 4, StanceSentences and StanceSentencesFeat are equally balanced, with a stratified number of examples per class to enhance the evaluation of StanceFit and StanceFeat' performance and explainability.

**Table 4**

*Datasheet for StanceSentences and StanceSentencesFeat Datasets for Stance Classification.*

| | Text Dataset | Feature Dataset |
|---|---|---|
| Name | *StanceSentences* | *StanceSentencesFeat* |
| Instances | Sentences from political discourses | Extracted linguistic features numerically represented |
| Classes (*) | • Support (*s*)<br>• Oppose (*o*) | • Support (*s*)<br>• Oppose (*o*) |
| Number of Instances | 1,280 (640 *s* / 640 *o*) | 1,280 (640 *s* / 640 *o*) |
| Instance Length | Between 5 to 50 tokens | 10 features |
| Labels | • "support"<br>• "oppose" | • "support"<br>• "oppose" |
| Splits/Instances | • Train: 972 (75.94%)<br>• Validation: 108 (8.44%)<br>• Test: 200 (15.62%) | • Train: 1080 (84.38%)<br>• Test: 200 (15.62%) |
| Stratification | • Train: 486 *s* and 486 *o*<br>• Validation: 54 *s* ad 54 *o*<br>• Test: 100 *s* and 100 *o* | • Train: 540 *s* / 540 *o*<br>• Test: 100 *s* / 100 *o* |
| Metadata | • title (document)<br>• source<br>• semantic_frames | • title (document)<br>• source<br>• semantic_frames |
| Data Period | 1939-2023 | 1939-2023 |

*Note:* (*) *s* = support, *o* = oppose. Both datasets are freely available in Hugging Face: (1) StanceSentences, DOI: 10.57967/hf/2652, and (2) GitHub, https://github.com/pacoreyes/stance_classification.

## 3.2 Experimental Set-up

We followed the choice of Biber and Finegan (1989) to use adjectives, adverbs, and verbs as features to analyze stance since it is aligned with established linguistic methodologies that use the role of particular *parts of speech* (*POS*) in conveying attitudes, evaluations, and orientations toward the content being discussed, and also due to the high granularity that offers this approach. (We return to this issue in bias mitigation.) Our feature engineering process based on linguistics began manually classifying each adjective, adverb, and verb found in the 1,280 sentences in the StanceSentences dataset into nine linguistic features (Table 5). We used previous studies, grammars, dictionaries, thesaurus, and querying ChatGPT (versions *4* and *4 Omni*) (2024) when disambiguation was necessary.

**Table 5**

*Linguistic Features used to Build the StanceSentencesFeat dataset.*

| Feature | Examples |
|---|---|
| **AFFECT** | |
| 1. Positive | |
| Adjectives | *fortunate*, *meaningful*, *transformative*, *top-of-the-line* |
| Adverbs | *successfully*, *democratically*, *tirelessly*, *mutually* |
| Verbs | *achieve*, *strive*, *immunize*, *empower* |
| 2. Negative | |
| Adjectives | *aggressive*, *disastrous*, *toxic*, *unpopular* |
| Adverbs | *unfortunately*, *negatively*, *overwhelmingly*, *arbitrarily* |
| Verbs | *aggravate*, *waste*, *endanger*, *misuse* |
| **EVIDENTIALITY** | |
| 3. Certainty | |
| Adjectives | *absolute*, *inevitable*, *conducive*, *well-known* |
| Adverbs | *obviously*, *of course*, *explicitly*, *therefore* |
| Verbs | *believe*, *establish*, *have shown*, *make sure* |
| 4. Emphatics | |
| Adjectives | *vigorous*, *groundbreaking*, *clear*, *eternal* |
| Adverbs | *incredibly*, *a lot*, *in particular*, *foremost* |
| Verbs | *focus*, *maintain*, *consolidate*, *make clear* |
| 5. Doubt | |
| Adjectives | *ambiguous*, *undetermined*, *distrustful*, *untrue* |
| Adverbs | *probably*, *perhaps*, *possibly*, *eventually* |
| Verbs | *think*, *expect*, *hope*, *attempt* |
| 6. Hedges | |
| Adjectives | *moderate*, *little*, *likely*, *several* |
| Adverbs | *maybe*, *alternatively*, *kind of*, *however* |
| Verbs | *lower*, *degrade*, *deter*, *hold down* |
| **MODALITY** | |
| 7. Modal verbs | |
| Predictive Modals | *will*, *would*, *shall*, *going to* |
| Possibility Modal | *can*, *may*, *might*, *could* |
| Necessity Modal | *ought to*, *should*, *must*, *need to*, *have to* |

| Feature | Examples |
|---|---|
| **POLARITY** | |
| 8. Pro polarity | |
| Adjectives | *committed*, *supportive*, *favorable*, *major commitment* |
| Adverbs | *together*, *on board* |
| Verbs | *bolster*, *favor*, *foster*, *make a commitment* |
| 9. Con polarity | |
| Adjectives | *opposed*, *unwilling*, *criticized*, *denounced* |
| Adverbs | *back*, *detrimentally*, *largely against*, *in opposition* |
| Verbs | *fight*, *disagree*, *resist*, *struggle against* |

We used the categories of stance features studied by Biber and Finegan (1989): *affect*, *evidentiality*, and *modality*. Affect is the range of expressed personal attitudes, which includes *positive* and *negative* emotions, feelings, moods, and general mental states (Ochs and Schieffelin, this issue). Conversely, evidentiality concerns how the speakers know the information they are discussing and their confidence in its accuracy (*epistemic certainty*, *epistemic doubt*, *emphatics*, and *hedges*). Modality refers to expressions that convey a speaker's attitude toward the *possibility*, *necessity*, or *predictive* of the state of affairs described by the verb in a clause. By empirical observation of subjective stance, we found that many sentences contained terms that expressed clear stance direction and did not fit in the other categories/features; for instance, actions that oppose ("*refute*", "*reject*", or "*contradict*") or that support ("*make a commitment*", "*advocate*", "*foster*"); therefore, we included in our analysis the features *pro polarity* and *con polarity*.

The classification of each term involved understanding its context of use in the stance expression, accepting only, in exceptional cases, a (*polysemous*) term in more than one feature to avoid the risk of multicollinearity in further data analysis stages. (We discuss that later in the feature engineering section.) This decision led us to make opinionated classifications, for instance, while "*obviously*" could be considered as both emphatics and certainty feature, we decided to put it in certainty, as that was its primary function in stance expressions within our corpus. Regarding emphatics, we accepted terms that generally are considered descriptive, but in our corpus had an empirical function to emphazise statements; for instance, "large" in "*The bottom line is, is that we think that Russia is a large important country with a military that is second only to ours, and has to be a part of the solution on the world stage, rather than part of the problem*." Also, we considered some determiners as adjectives; for instance, "*any*" in "*I can only say with emphasis, we vigorously oppose any government in NATO that would have a Communist head or control--vigorously.*" We rejected neutral terms that did not help statements to define a position ("*act*", "*necessitate*", or "*afford*"), ambiguous terms that did not contributed to understand the stance expression ("*protectionist*", "*strategic*", or "*meet*") or those that may suppose ideological or thematic adhesion ("*liberal*", "*military*" and "*united*"—most of the time from "*United States*"), or when nouns were used as adjectives ("*tax*" in "*tax legislation*" or "*education*" in "*education reform*"). However, we accepted terms that in other domains may be neutral or merely descriptive but in our particular use case were positional, such as "*democratic*" as a positive adjective, or "*nuclearize*" as a negative verb, since they had a clear affective connotation in American politics. Also, we accepted phrases that showed a clear function in stance, for instance: the pattern "*real* [adjective]" was considered as an emphatics adjective, the

expression "*sort of*" as a hedges adverb, "*recognize importance*" as a pro polarity verb, or "*strengthen the opposition*" as a con polarity verb. Also, we removed demonyms, locationals ("*national*" or "*international*"), named-entity terms and any other feature with low representativity or that jeopardized the thematic neutrality to our dataset. Finally, we made more arbitrary classifications under a deep understanding of our domain; for instance, although the verb "*believe*" is usually not considered to convey certainty in linguistics literature, we found it in our domain used in certainty expression; similarly, we added the phrase "*believe in*" as a pro polarity verb. The list of studied features can be found in the Appendix.

This feature engineering modelled the LRBM combining two approaches: (1) *lexicon-based* matching, using dictionaries of adjectives, adverbs, and verbs, and (2) *pattern-based* matching, using spaCy's linguistic feature analysis capacities, such as tokenization, lemmatization, POS tagging, dependency parsing, and pattern matchers (*Matcher* and *Phrasematcher*). The LRBM detected, evaluated, and scored each feature as a single-token term ("*assure*") or multi-token term ("*well-known*"). The matchers verified the POS of each term; for instance, the adjective "*well*" was distinguished from the adverb "*well*". This process generated a binary list (*0* or *1*) representing each token in a sentence, excluding punctuation. A score of *1* was assigned if one or more tokens matched a term in the lexicon or a specified pattern. Tokens that did not match any rule received a score of *0*. The output conformed to the StanceSentencesFeat dataset, comprising numeric representations of linguistic features extracted from the StanceSentences dataset.

During the initial *exploratory data analysis* (*EDA*), we observed two issues that defined feature aggregation and scoring:

1. *Sparse data*: The prevalent absence of most features in sentences created high negative skewness; thus, we opted for binarization, scoring *1* if a feature was present in the sentence and *0* if not; and
2. *Feature granularity*: Following the feature aggregation in Biber and Finegan (1989)—in which emphatics was aggregated to certainty, and hedges was aggregated to doubt—the LR analysis reported low statistical significance for certainty and doubt; thus, we opted to treat emphatics and hedges as independent features, reducing *loss information* and increasing *dimensionality* in the LRBM to capture the complexity of political language with more granularity.

*The feature aggregation* resulted in modeling the LRBM with eight linguistic features:

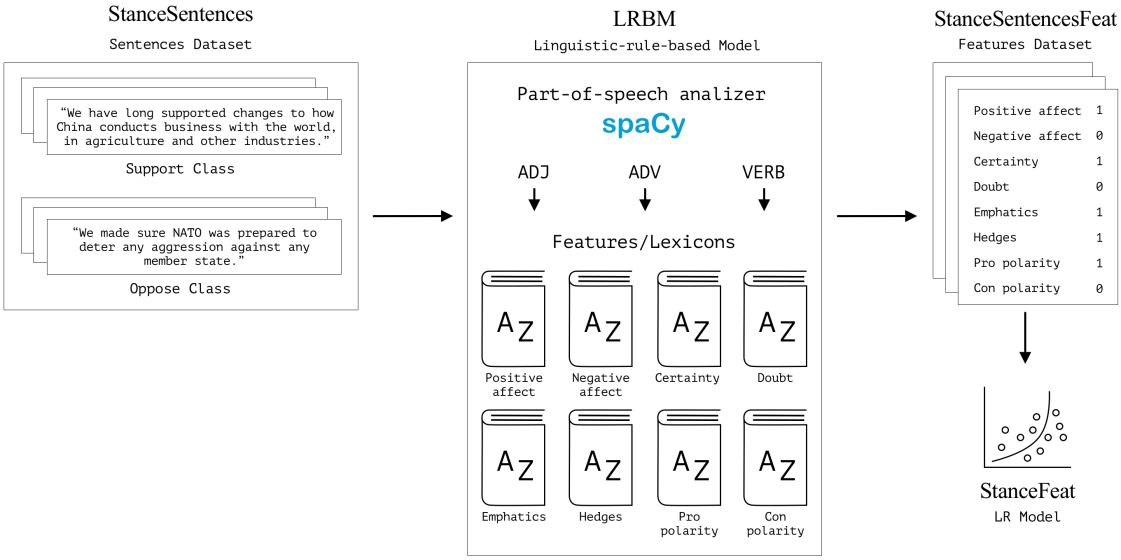
1. *Positive affect*: The count of positive adjectives, adverbs, and verbs.
2. *Negative affect*: The count of negative adjectives, adverbs, and verbs.
3. *Certainty*: The count of certainty adjectives, adverbs, and verbs.
4. *Doubt*: The count of doubt adjectives, adverbs, and verbs.
5. *Emphatics*: The count of emphatics adjectives, adverbs and verbs, and predictive modal verbs.
6. *Hedges*: The count of hedges adjectives, adverbs and verbs, possibility modal verbs, and necessity modal verbs.

7. *Pro polarity*: The count of pro adjectives, adverbs, and verbs.
8. *Con polarity*: The count of con adjectives, adverbs, and verbs.

After LR analysis, we pruned features with low significance, and after iterations to evaluate the best features to retain, we fit StanceFeat with StanceSentencesFeat, using the training split (80%) and the test split (20%) (Figure 2). The LR analysis of StanceFeat paved the way for a deeper investigation into StanceFit's explainability and ability to generalize.

**Figure 2**

*Process to use the Linguistic-Rule-Based Model (LRBM) to Extract Numerical Representations of Linguistic Features from Stance Expressions.*




We chose SetFit's special variation *paraphrase-mpnet-base-v2* to create StanceFit since it aligns with the challenges of classifying political discourse texts since it is especially adept at discerning the subtle differences in sentences that may convey similar messages with different wording. We used the StanceSentences dataset, divided into training (80%), validation (10%), and testing (10%) subsets, to fit StanceFit on a *Google Colab* environment with an NVIDIA A100 GPU with 40GB of VRAM and the PyTorch deep learning framework (Paszke et al., 2019). We used *Optuna* (Akiba et al., 2019) to find the best model by maximizing accuracy. We monitored the *training* and *evaluation embeddings* closely by saving *t-SNE* (*t-distributed Stochastic Neighbor Embedding*) plots to follow up overfitting. We used the following hyperparameters in iteration 10 of bootstrapping: *body learning rate*, 1.003444469523018e-06; *batch Size*, 16; *max iterations*, 237; *number of epochs*, 3; *solver*, lbfgs; and *seed*, 37. We used Python's libraries for data manipulation and visualization, such as *Pandas* (The pandas development team, 2020), *Seaborn* (Waskom, 2021), *Matplotlib* (Hunter, 2007), and *Scikit-learn* (Pedregosa et al., 2011).

We selected SHAP as an explainability tool due to its ability to elucidate the contribution of each token in StanceFit's decision-making process, enhancing our understanding of neural networks' interpretability in stance classification. Consequently, our methodology shifted to token-level analysis, deactivating spaCy's matchers for phrase constructions and multi-token patterns in the LRBM. This change was validated using LR analysis, which indicated a negligible impact on overall results due to the low frequency of multi-token terms.

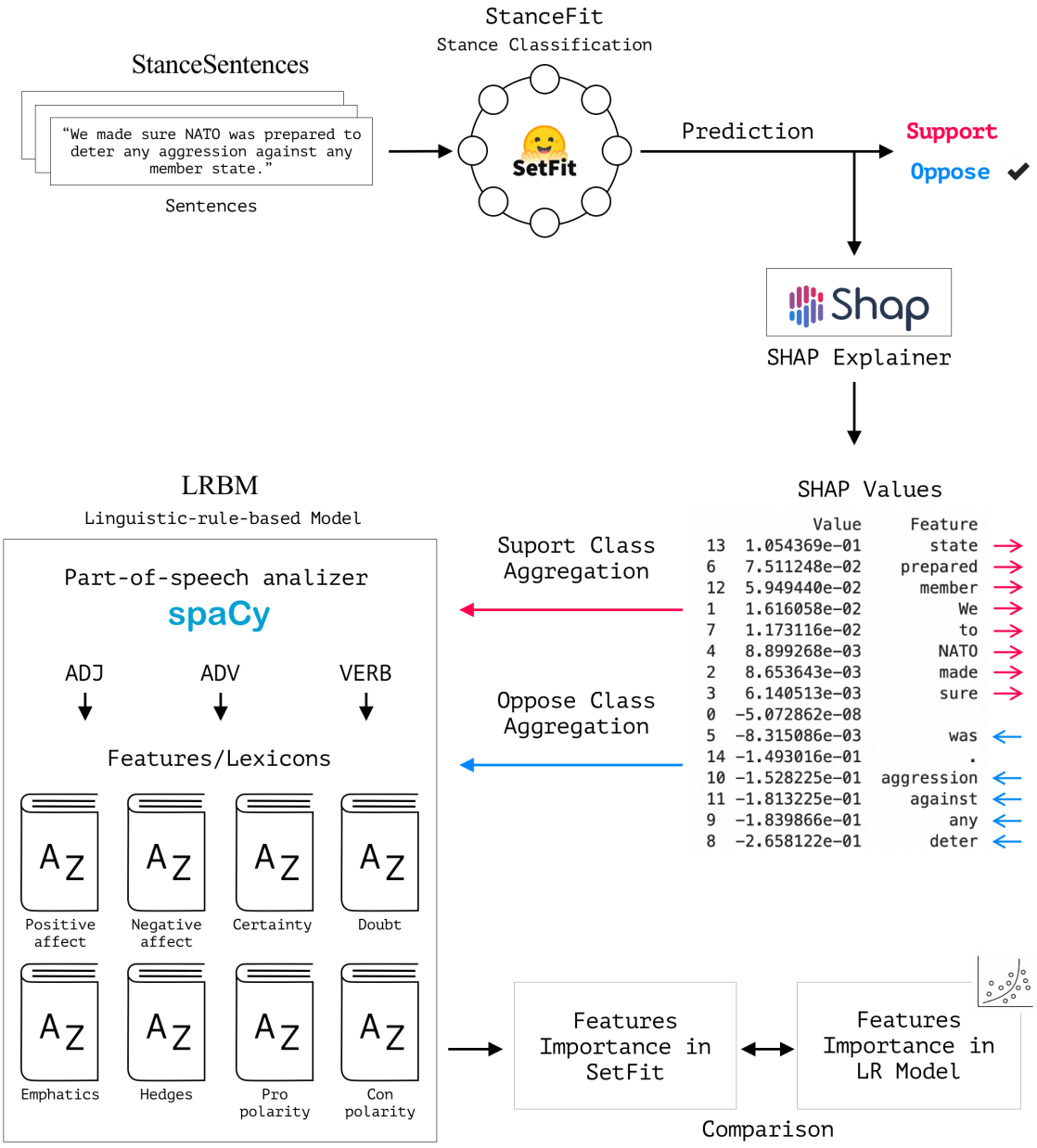
To compare the influence of eight linguistic features across both models, we calculated aggregated SHAP values per feature, suitable for comparison with LR coefficients. We applied SHAP on the test and validation splits of StanceSentences, unseen by StanceFit. Aggregated mean SHAP values were derived by summing positive and negative mean SHAP values for a feature whenever LRBM identified a term in a lexicon and confirmed its POS in a sentence. The aggregated SHAP value for feature $i$ ($SHAP_i$) is defined as:

$$SHAP_i = \sum_{j=1}^{n} \text{mean}(SHAP_{ij})$$

where $SHAP_{ij}$ represents the SHAP value for feature $i$ in observation $j$, and $n$ is the total number of observations. We normalized values from both models using scikit-learn's StandardScaler (which handles numerical signs, useful to retain stance directionality) to facilitate comparison with LR coefficients. Figure 3 illustrates the NLP explainability process using SHAP.

**Figure 3**

*NLP Explanability Process Using SHAP on StanceFit, the SetFit model.*



Finally, a two-sample (*Welch*) t-test compared LR coefficients and SHAP values, and to give a more visual evaluation of the agreement of the compared features for each model, we ran a Bland-Altman analysis to confirm if both models (StanceFeat and StanceFit) assign similar importance to the linguistic features in classifying stance in political discourse.

### 3.3 Bias Mitigation

Bias mitigation efforts commenced during the feature selection process. We controlled thematic bias by focusing on specific POS (adjectives, adverbs, and verbs), typically conveyed through nouns—especially named entities. This control extended to the lexicon built process, where we excluded or reduced denominal adjectives ("*war*" in "*war taxes*" ), adverbs ("*legislatively*" in "*legislatively, this bill aims*), and verbs ("*institutionalize*" in "*institutionalizing free and open*") to prevent thematic adherence.

Further bias mitigation occurred during the dataset annotation process. We controlled the presence of the verb and noun "*war*" in examples, being prevalent in both classes but dominant in the oppose class. Since war is present in many moments of American history, politicians often take a position on the "*war*" issue in political discourse. Also, the "*war*" word, alongside many other belligerent and militaristic synonyms ("*combat*", "*struggle*", "*fight*", "*battle*", Etc.), is used to describe oppose stance toward other unrelated issues: "*combat climate change*", "*war on terror*", "*struggle against disease*", "*fight corruption*", "*battle Coronavirus*", or "*long-range*", a military synonym of "*long-term*" or "*future*" ("*I believe it is a good investment in momentum and a long-range possibility of an equitable and secure peace in the Middle East*."). The same issue happened in the oppose class, where political issues like "*trade*" or "*education*" were more prevalent and needed mitigation. Therefore, we consciously handled certain terms whenever possible.

Generally, we noted a scarcity of representative examples for the support class, where stance is more frequently expressed with ambiguity, reflecting how politicians often frame their discourse to emphasize agreement, collaboration, and positive action, even when addressing contentious issues. This tendency necessitated additional efforts to identify examples of direct subjective stance expressions for the oppose class.

## 4. Results and Discussion

### 4.1 Feature Engineering

We can draw several conclusions from the distribution of linguistic features across both stance classes (Table 6). Political statements indicating support frequently employ more positive affect (support: 648; oppose: 201) and pro polarity terms (support: 523; oppose: 167), emphasizing favorable and assertive language. Conversely, statements expressing opposition exhibit higher usage of negative affect (support: 53; oppose: 315) and con polarity terms (support: 77; oppose: 510), reflecting a critical tone. However, the total of negative affect terms in the oppose class is half the total of positive affect terms in the support class (support: 648; oppose: 315). Similarly, the count of pro polarity verbs in the oppose class is 141, while the con polarity verbs in the support class are only 62. This disproportion in both cases suggests that politicians often frame their discourse with apparent agreement when they are being confrontational. Certainty and emphatics are notably prevalent in both support (certainty: 434; emphatics: 1,068) and oppose (certainty:

383; emphatics: 997) stances, suggesting a strong assertiveness across both classes. Although hedges is generally low in both stance classes (support: 78; oppose: 123)—in line with the intentional selection of direct stance expressions—there is a clear inclination towards the oppose class, indicating the already observed need to soften opposing statements. Modality features such as necessity (support: 236; oppose: 292) and possibility (support: 62; oppose: 54) further delineate the commitment and hypothetical scenarios often invoked in political statements.

**Table 6**

*Distribution of the Count of terms classified by Feature and POS in dataset StanceSentences for Stance Classification in Support and Oppose Classes.*

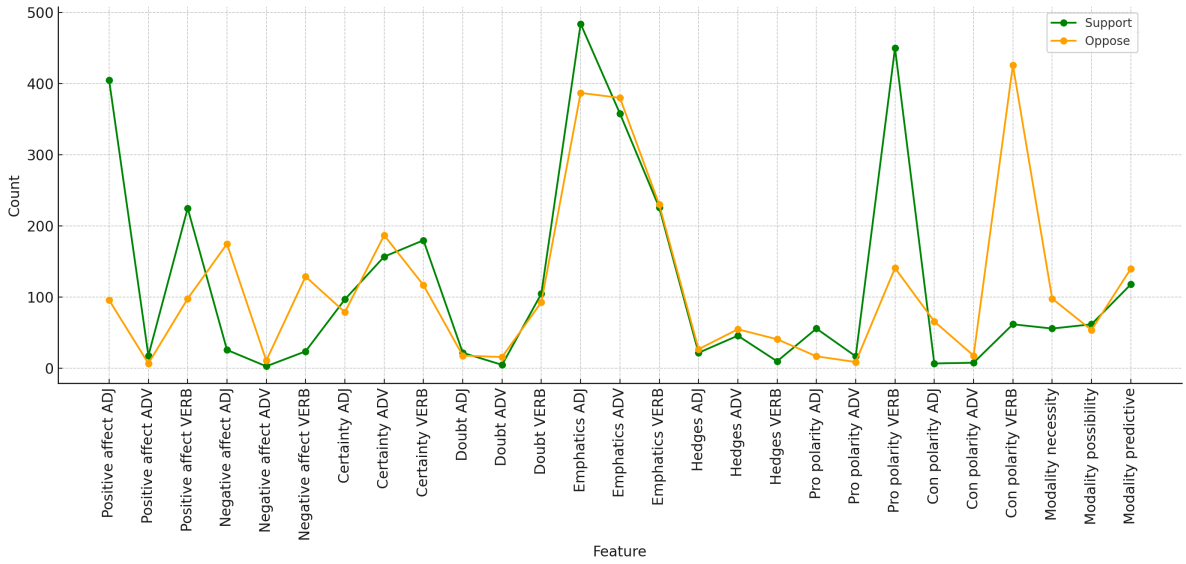
| Feature | *POS* | *Support* | | *Oppose* | | Total |
|---|---|---|---|---|---|---|
| | | Count | Total | Count | Total | |
| Positive affect | ADJ | 404 | **648** | 95 | **201** | 847 |
| | ADV | 18 | | 7 | | |
| | VERB | 225 | | 98 | | |
| Negative affect | ADJ | 26 | **53** | 175 | **315** | 368 |
| | ADV | 3 | | 11 | | |
| | VERB | 24 | | 129 | | |
| Certainty | ADJ | 97 | **434** | 79 | **383** | 817 |
| | ADV | 157 | | 187 | | |
| | VERB | 180 | | 117 | | |
| Doubt | ADJ | 22 | 132 | 18 | 127 | 259 |
| | ADV | 5 | | 16 | | |
| | VERB | 105 | | 93 | | |
| Emphatics | ADJ | **484** | **1,068** | **387** | **997** | 2,065 |
| | ADV | 358 | | 380 | | |
| | VERB | 226 | | 230 | | |
| Hedges | ADJ | 22 | **78** | 27 | **123** | 201 |
| | ADV | 46 | | 55 | | |
| | VERB | 10 | | 41 | | |
| Pro polarity | ADJ | 56 | **523** | 17 | **167** | 690 |
| | ADV | 17 | | 9 | | |
| | VERB | 450 | | **141** | | |
| Con polarity | ADJ | 7 | **77** | 64 | **510** | 587 |
| | ADV | 8 | | 18 | | |
| | VERB | **62** | | 426 | | |
| Modality | Necessity | 56 | **236** | 98 | **292** | 528 |
| | Possibility | 62 | | 54 | | |
| | Predictive | 118 | | 140 | | |

Figure 4 visually illustrates distinct trends in the distribution of linguistic features across both stance classes. Notably, a clear correspondence between the count of positive affect terms exhibited a strong association with pro polarity. Conversely, the count of negative affect terms aligned significantly with con polarity, suggesting expressions of positive

emotion are more frequently employed while expressing support stance, and expressions of negative emotion are predominantly utilized in oppose stance. This finding underscores the alignment between affective language and stance, reflecting the emotional undertones embedded within pro and con arguments.

**Figure 4**

*Distribution of Linguistic Features by POS in Support and Oppose Classes for Stance Detection in StanceSentences dataset.*




The LR analysis (Table 7) provided insights into the importance of features, with pro polarity and positive affect being the most important predictors towards the support class, having pro polarity a coefficient ($\beta$) of 1.867 and odds ratio ($e^{\wedge}\beta$) of 6.468, and positive affect with a $\beta$ of 1.777 and an $e^{\wedge}\beta$ of 5.914, both having high $z$-values (9.772 and 1.777, respectively) and very low $p$-values ($p < .001$), indicating that the presence of words expressing pro polarity or positive affect increases the likelihood of a sentence expressing a supportive stance. Oppositely, negative affect and con polarity are significant predictors towards the oppose class, with negative affect with a $\beta$ of -1.980 and $e^{\wedge}\beta$ of 0.138, and con polarity with a $\beta$ of -2.723 and an $e^{\wedge}\beta$ of 0.065, both having negative coefficients and very low $p$-values, suggesting that sentences with these features are most probably expressions of opposing stance.

**Table 7**

*LR Results for the Classification of Support and Oppose Stance Features.*

| Feature | $\beta$ | $SE\beta$ | $z$ | $p$ | $e^{\wedge}\beta$ |
|---|---|---|---|---|---|
| const (intercept) | -0.314 | 0.274 | -1.144 | 0.252 | 0.731 |
| Pro polarity | **1.867** | 0.191 | **9.772** | 0.000 | **6.468** |
| Positive affect | **1.777** | 0.184 | **9.641** | 0.000 | **5.914** |
| Doubt | 0.309 | 0.222 | 1.393 | **0.164** | 1.362 |
| Emphatics | 0.033 | 0.232 | 0.139 | **0.889** | 1.033 |
| Certainty | 0.138 | 0.177 | 0.780 | **0.435** | 1.148 |

| Feature | *β* | *SEβ* | *z* | *p* | *e^β* |
|---|---|---|---|---|---|
| Hedges | **-0.321** | 0.193 | -1.658 | **0.097** | 0.726 |
| Negative affect | **-1.980** | 0.239 | -8.284 | 0.000 | **0.138** |
| Con polarity | **-2.723** | 0.203 | -13.417 | 0.000 | **0.065** |

Certainty, emphatics, and doubt did not show a significant impact on stance classification; certainty with a $\beta$ of 0.138 and a $p$-value of 0.435; emphatics with a $\beta$ of 0.033, and a $p$-value of .889; and doubt with a $\beta$ of 0.309 and a $p$-value of 0.164; all of them had $p$-values above the 0.05 threshold, indicating that their influence on stance detection is statistically insignificant in this model. Finally, hedges reported a $\beta$ of -0.321 and a $p$-value of 0.097, suggesting that, albeit weak, it is associated with expressions of opposing stance. This finding aligns well with an observed similar pattern in the analysis of distribution of features lines above. After an iterative feature optimization process evaluating the accuracy of the StanceFeat, we decided to retain doubt, and prune certainty and emphatics.

## 4.2 Bootstrapping Process

Table 4 shows the metrics of StanceFit along the bootstrapping process to build the StanceSentences dataset.

**Table 8**

*Summary of Performance Metrics of StanceFit during the Bootstrapping Process to Build the StanceSentences dataset.*

| Iteration | Dataset length | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 (seed) | 180 | 0.960 | 0.961 | 0.960 | 0.960 |
| 2 | 280 | 0.965 | 0.965 | 0.965 | 0.965 |
| 3 | 380 | 0.970 | 0.971 | 0.970 | 0.970 |
| 4 | 480 | 0.975 | 0.975 | 0.975 | 0.975 |
| 5 | 580 | 0.980 | 0.980 | 0.980 | 0.980 |
| 6 | 680 | 0.980 | 0.980 | 0.980 | 0.980 |
| 7 | 780 | 0.985 | 0.985 | 0.985 | 0.985 |
| 8 | 880 | 0.990 | 0.990 | 0.990 | 0.990 |
| 9 | 980 | **0.995** | 0.995 | 0.995 | 0.995 |
| 10 | 1,080 | **0.995** | 0.995 | 0.995 | 0.995 |

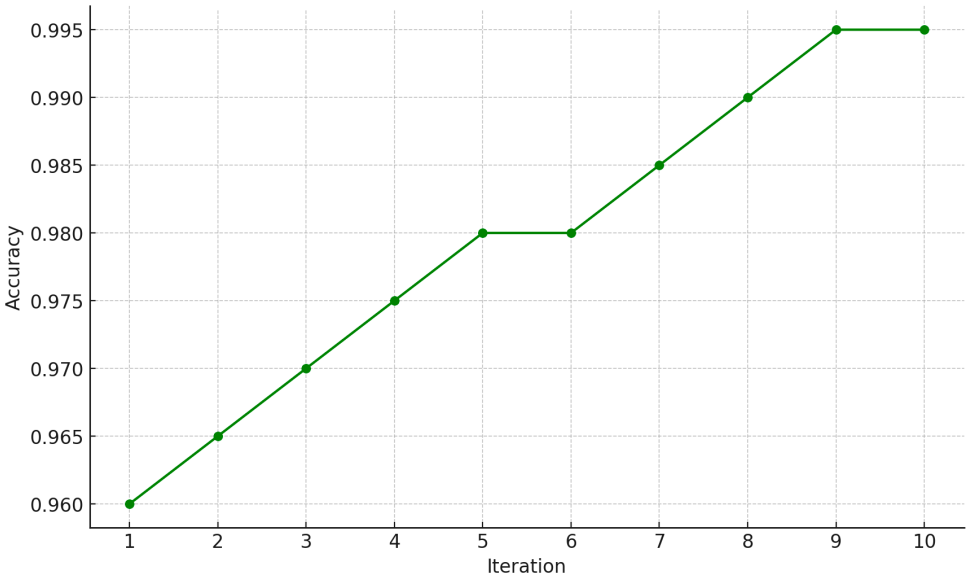
*Note:* In each iteration, the model was tested using the same test dataset of 200 examples.

The general trend observed was increased performance as more data was incorporated into training through bootstrapping, indicating that SetFit's capacity to leverage incremental data effectively enhances its predictive accuracy. Such a pattern is typical in few-shot learning scenarios where initial training data is limited, and each additional bit of data can significantly refine the model's understanding (Gao et al. 2021). As more data was added, the rate of improvement in model performance tends to plateau, a phenomenon clearly visible in the latter iterations, suggesting that the model extracted as much generalizable knowledge as

it can from the data provided. In general, as observed in Figure 5, the data analysis suggests that StanceFit adapted well to incremental data in a bootstrap training framework.

**Figure 5**

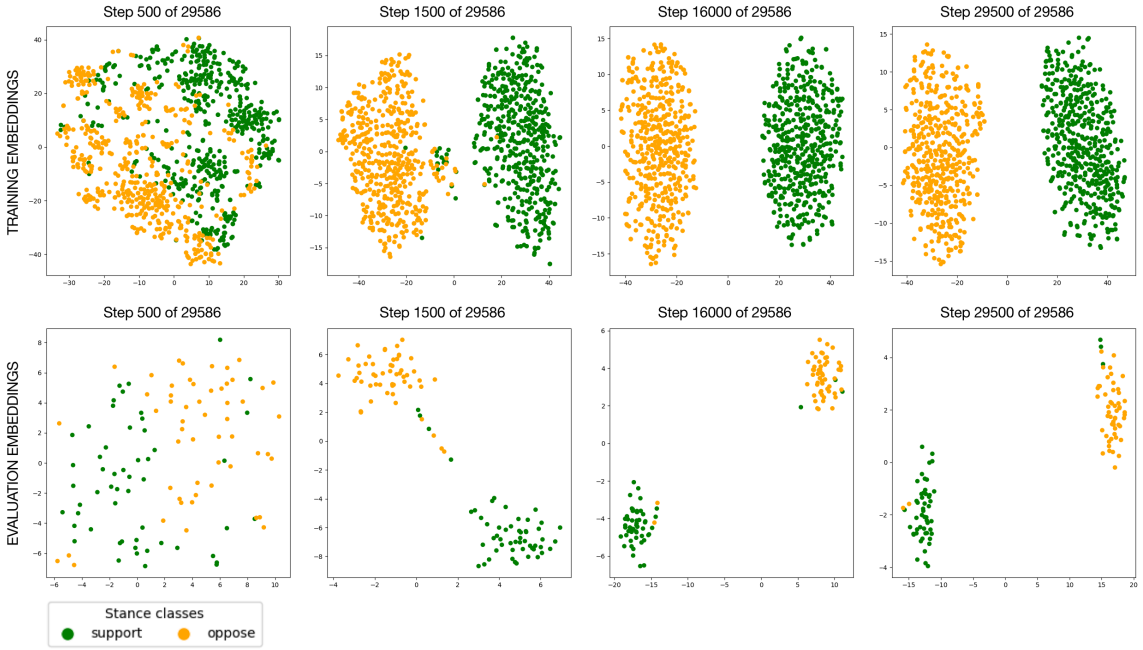
*Accuracy of the StanceFit Across Iterations of Bootstrapping Training*.




To visually evaluate the adaptation of the model to the classification task, Figure 6 shows the t-SNE plots generated along the fine-tuning process of iteration 10 where StanceFit is adapting its embeddings to the classification task. The training embeddings display how sentences have been converted to dense vector spaces where semantically similar sentences are placed closely together and dissimilar ones are distant, confirming that StanceFit discerns and groups stance expressions, distinguishing nuanced differences between both stance classes: support (green) and oppose (orange). The sequence of visualizations depicts how StanceFit evolved over time for stance classification on the distinct embedding separations. In the early training stage, step 500, the embeddings of both classes are mixed together, indicating that the model has not yet learned to differentiate between the two stances effectively. However, soon after, in step 1,500, some basic structure starts to form where clusters begin to separate slightly, though some overlap remains. By step 16,000, the clusters are well-separated, suggesting that the model has learned distinctive features for each class. Finally, in step 29,500, the clusters are distinctly separate with minimal overlap, implying a high level of learning and specialization in distinguishing between the two stances.

**Figure 6**

*t-SNE plots of Iteration 10 of the Bootstrapping Training for Stance Classification*.




The evaluation embeddings validate how well the model fit to the learned data, showing that the Sentence Transformers architecture of the model successfully maps the *evaluation* sentences into a space where their stances can be accurately classified, even when the model has not directly learned from these specific examples. Despite different training stages, the consistent separation in *evaluation embeddings*—including some misclassifications—indicates that the model is not just memorizing the training data but understanding the features that define each class; misclassifications in evaluation data are expected and normal.

## 4.3 Models Performance

As seen in Table 9, StanceFit achieves near-perfect scores (0.995 or higher) in all metrics, indicating its high effectiveness in stance classification at the sentence level. In contrast, the baseline model, StanceFeat (LR model), shows good performance but is considerably lower than StanceFit. Specifically, StanceFeat, with an AUC-ROC of 0.815, demonstrates a reasonably good ability to differentiate between classes. However, the StanceFit model achieves an exceptionally high AUC-ROC value of 0.995, suggesting that StanceFit almost always correctly classifies the stance.

**Table 9**

*Summary of Performance Metrics of StanceFeat and StanceFit for Classifying Support and Oppose Stance at the Sentence Level.*

| Metric | LR | SetFit |
|---|---|---|
| Name | **StanceFeat** | **StanceFit** |
| Accuracy | **0.810** | **0.995** |
| Precision (macro) | 0.823 | 0.995 |
| Recall (macro) | 0.790 | 0.995 |
| F1 Score (macro) | 0.810 | 0.995 |
| AUC-ROC | **0.815** | **0.995** |
| Confusion Matrix (*) | *s o* | *s o* |

| Metric | LR | | SetFit | |
|---|---|---|---|---|
| | *s* | 84 17 | *s* | 100 0 |
| | *o* | 21 79 | *o* | 1 99 |
| Support Class | | | | |
| Precision | 0.798 | | 0.990 | |
| Recall | 0.830 | | 1.000 | |
| F1-score | 0.814 | | 0.995 | |
| Oppose Class | | | | |
| Precision | 0.823 | | 1.000 | |
| Recall | 0.790 | | 0.990 | |
| F1-score | 0.806 | | 0.995 | |

*Note:* (*) *s* = support, *o* = oppose. (1) Across-class metrics are macro and class-wise metrics are not averaged. (2) StanceFit, DOI: 10.57967/hf/2618, is freely available in Hugging Face.

The confusion matrices show that both models struggle with the oppose class, suggesting that expressions of opposition or disagreement in political discourse are more challenging to understand and classify. Expressions with opposing stances might use negations, conditionals, or other more complex syntactic structures that implicitly convey disagreement without expressing it explicitly. Usually, the vocabulary used in opposing statements may vary widely, generating a variability that could make it more difficult for models, notably simpler ones like LR, to capture and generalize across different expressions of opposition. The superior performance of StanceFit can be attributed to its unique technology, embeddings in Sentence Transformers. This technology enables the model to handle nuanced language, contextual understanding, irony, sarcasm, emotional overtones, and ambiguity, making it more effective to understand language at the semantics and pragmatics levels.

The use of the special version of SetFit, paraphrase-mpnet-base-v2, further enhances its capabilities, as it leverages the neural network architecture *MPNet*, which combines the strengths of both masked language modeling (BERT) and permuted language modeling (XLNet). This unique combination allowed paraphrase-mpnet-base-v2 to detect stances by capturing linguistic cues that might have not be explicitly clear but are implied through paraphrasing or similar expressions.

## 4.6 SHAP Analysis

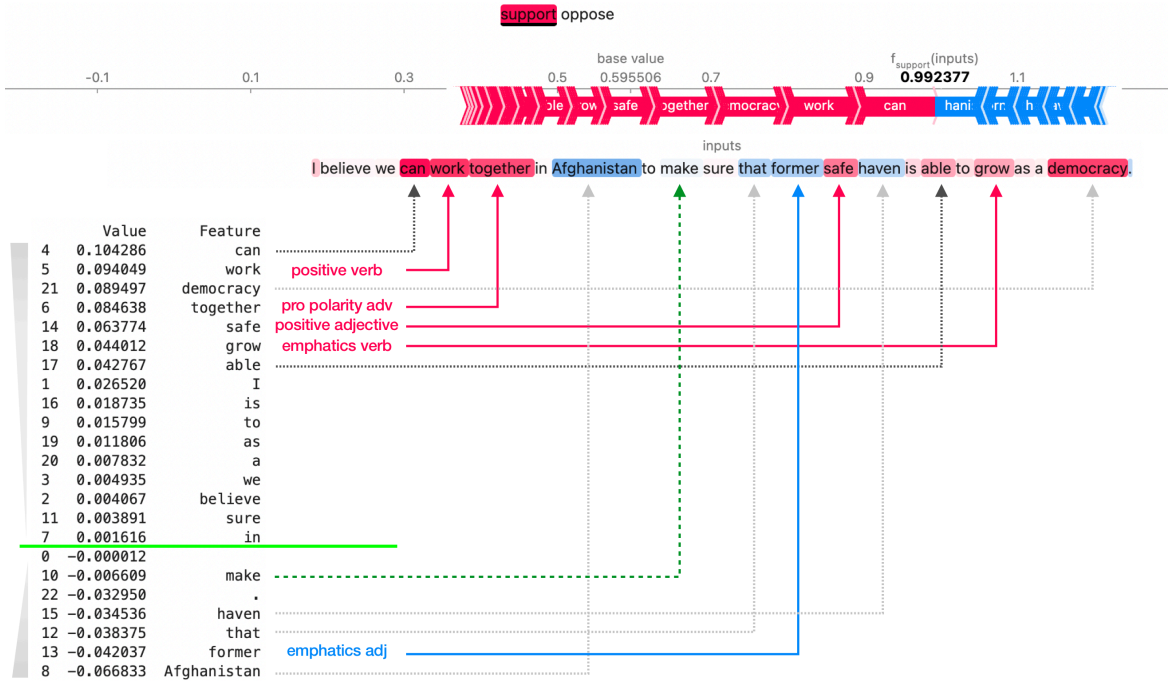
The following two examples depict our linguistic-aware approach to explain inference behaviors from StanceFit by comparing SHAP values and prediction coefficients of the linguistic-aware LR, StanceFeat, model through LRBM.

In Example 1 (Figure 7), the SHAP's *text plot* visualization shows how individual tokens in a sentence influence StanceFit's decision-making process, classified as a support stance expression (the support label on the top with red color). In red are the tokens that contribute positively towards the prediction of the support class (positive SHAP values), and in blue are the tokens that contribute positively towards the prediction of the oppose class (negative SHAP values). The intensity of the color corresponds to the magnitude of the token's

contribution, with deeper shades indicating stronger contributions. The list of tokens (features) with their respective SHAP values is at the bottom. The horizontal green line divides the list into positive SHAP values (ascending values for support) above the line and negative SHAP values (descending values for oppose) below the line, such that the highest values for each class are at the respective ends of the ranking.

**Figure 7**

*Example 1: SHAP explanation on an example of the Support Class and Alignment to Lexicons in the LRBM.*




| | Value | Feature |
|---|---|---|
| 4 | 0.104286 | can |
| 5 | 0.094049 | work |
| 21 | 0.089497 | democracy |
| 6 | 0.084638 | together |
| 14 | 0.063774 | safe |
| 18 | 0.044012 | grow |
| 17 | 0.042767 | able |
| 1 | 0.026520 | I |
| 16 | 0.018735 | is |
| 9 | 0.015799 | to |
| 19 | 0.011806 | as |
| 20 | 0.007832 | a |
| 3 | 0.004935 | we |
| 2 | 0.004067 | believe |
| 11 | 0.003891 | sure |
| 7 | 0.001616 | in |
| 0 | −0.000012 | |
| 10 | −0.006609 | make |
| 22 | −0.032950 | . |
| 15 | −0.034536 | haven |
| 12 | −0.038375 | that |
| 13 | −0.042037 | former |
| 8 | −0.066833 | Afghanistan |

This analysis follows the colors of the arrow lines in Figure 7:

1. *Support term in lexicons (red)*: The verb "*work*", the adverb "*together*", the adjective "*safe*", and the verb "*grow*" have higher values that contributed to the classification in the support class and were found in the lexicons.
2. *Oppose term in lexicons (blue)*: The adjective "*former*" had the higher value that contributed to the classification in the oppose class and was found in the lexicons.
3. *Omitted term in lexicons (dark gray)*: The verb "*can*" and the adjective "*able*" played a significant role in the classification of the support class, but none of them were found included in the lexicons. Two reasons for this omission help us to understand the limitations of the LRBM, but also its strengths: (1) we considered the verb "*can*" to have a neutral meaning, and therefore we excluded it from our analysis, and (2) the adjective "*able*" was disregarded during the lexicon building stage as a positive adjective—but in retrospective, this exclusion deserves to be corrected and classified as a positive adjective. On the other hand, StanceFit included the neutral verb "*can*" along with the other tokens, forming the phrase "*can work together*", meaning StanceFit made a sophisticated evaluation of stance including the context at the phrase level.
4. *Excluded POS (light gray)*: Although our analysis does not focus on nouns, we can investigate StanceFit's behavior on nouns. (1) "*Afghanistan*" (influencing towards the oppose class) and "*democracy*" (influencing towards the support class). "*Afghanistan*"

could have been added as a feature of the oppose class because of the belligerent language of American politics that associates opposition to certain themes and named entities. Something similar and in the opposite direction could have happened with the noun "*democracy*", which usually appears as a positive concept in the public sphere. (We return to this issue in the conclusion.) (2) the conjunction "*that*" and the noun "*haven*" influenced the classification toward the oppose class without any apparent reason, and it should be investigated.

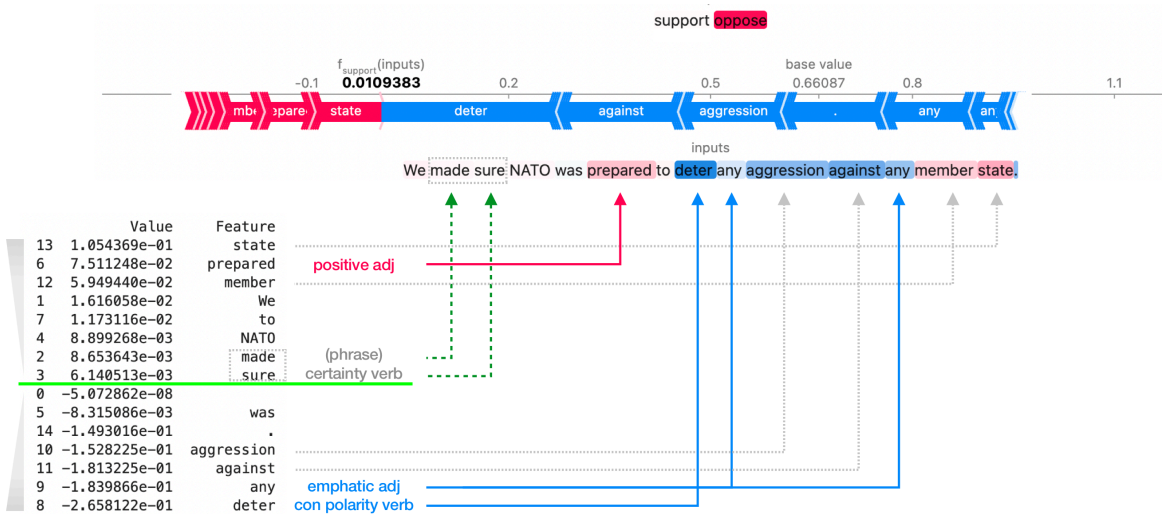
5. *Low-scored certainty (green dotted)*: The phrase "*make sure*", included as a certainty verb (see Appendix, 9. Certainty verbs) received a low score by SHAP, which aligns with the low LR coefficient of the certainty feature reported in Table 7.

Overall, this analysis reports that positive affect and pro-polarity features had a strong influence on the classification of stance toward the support class, as observable in the ranking of tokens with red color. On the contrary, the token "*grow*", classified as an emphatic verb, was graded lower. An apparent counter-intuitive behavior of the model is the low score given to the phrase "*believe in*" (see Appendix, 21. Pro verbs), which expresses a clear favorable stance. This behavior allows us to see the high adaptability of the model to this classification task, which overrides the influence of "*believe in*" due to the finding of features with a stronger influence on the support class. Finally, we observe that both nouns ranked higher, "*Afghanistan*" and "*democracy*", meaning that the adherence to topics—a concerning bias—is significant, but not conclusive due to the capacity of StanceFit to discriminate features and adapt to the classification challenge with flexibility.

In Example 2 (Figure 8), the SHAP's text plot visualization shows the analysis of a sentence classified by StanceFit as an oppose stance expression (the oppose label on the top with red color).

**Figure 8**

*Example 2: SHAP explanation on an example of the Oppose Class and Alignment to Lexicons in the LRBM.*




Through this example, we can observe other complementary behaviors in our analysis of StanceFit:

1. *Support term in lexicons (red)*: The adjective "*prepared*" has the higher SHAP value that contributed to the classification of the support class and was also found in the lexicons.
2. *Oppose term in lexicons (blue*): The verb "*deter*" and the adjective "*any*" (actually a determiner featured as an emphatics adjective) had the higher value that contributed to the classification of the oppose class and were also found in the lexicons.
3. *Excluded POS (light gray)*: The phrase "*aggression against*", composed of two POS excluded from our lexical analysis (noun and preposition), plays a significant, influential role in the oppose class, allowing us to observe—again—the capacity of the model to analyze sequences of tokens and phrases.
4. *Low-scored certainty (green dotted)*: The phrase "*made sure*", included (as its lemma form *make sure*) as a certainty verb (see Appendix, 9. Certainty verbs) received a low score by SHAP, which aligns with the low influence of the certainty feature which aligns with the low LR coefficient of the certainty feature reported in Table 7.

The second example confirms previous findings and introduces new ones. For instance, the stronger influence of features that signal the direction of the stance, in this case, con polarity in the verb "*deter*" that prevailed over other features to predict the oppose class. Towards the support class, we observed a high value in the adjective "*prepared*"; however, its value is low compared to those leading the oppose class. In this example, the strong influence of verbs ("*deter*") in the classification over contextual information provided via nouns ("*aggression*") is noticeable. The adaptation of StanceFit to recognize the verb "*deter*" as an oppose indicator is intriguing (probably attributed to the nature of its variation paraphrase-mpnet-base-v2), especially if we consider that "*deter*" is present only in two examples in the training dataset (both annotated as oppose). This finding speaks about the versatility of paraphrase-mpnet-base-v2. Finally, regarding nouns, the fact that "*NATO*" had a very low SHAP value in this example of the oppose class makes us think that its prevalence in the support class, makesStanceFot consider it irrelevant in the oppose class.

The comparative analysis between StanceFeat's LR coefficients ($\beta$) and StanceFit's SHAP values (Table 10) revealed alignments and discrepancies in the magnitude of features. Pro polarity exhibits a strong positive influence in both models, with normalized $\beta$ at 1.329 and SHAP value at 1.454, signaling its significant contribution to the support class. Positive affect also contributes positively, though to a lesser extent, with normalized values of 1.270 for $\beta$ and 0.930 as SHAP value. Negative affect and con polarity show significant negative contributions, with negative affect having normalized values of -1.235 $\beta$ and -0.827 as SHAP value, and con polarity exhibiting the strongest negative influence with values of -1.730 $\beta$ and -2.011 as SHAP value. These negative contributions are crucial as they influence the oppose class predictions. Certainty exhibits a modest positive effect, $\beta$ at 0.177 and SHAP value at 0.301, indicating a consistent but slight contribution to the support class. Doubt shows a small negative impact, $\beta$ at 0.210 and SHAP values at -0.172, highlighting its role in decreasing the likelihood of support.

**Table 10**

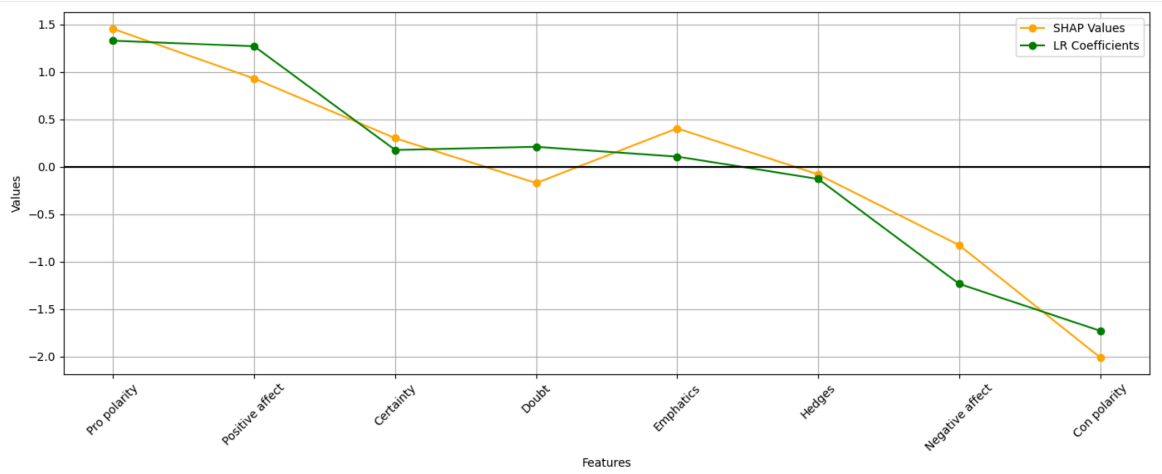
*Comparison of Feature Importance in StanceFeat's LR Coefficients and StanceFit's SHAP values.*

| Feature | Raw | | Normalized | |
|---|---|---|---|---|
| | LR $\beta$ | SHAP Values | LR $\beta$ | SHAP Values |
| Pro polarity | 1.867 | 12.776 | **1.329** | **1.454** |
| Positive affect | 1.777 | 8.258 | 1.270 | 0.930 |
| Certainty | 0.138 | 2.835 | 0.177 | 0.301 |
| Doubt | 0.309 | -1.240 | 0.210 | -0.172 |
| Emphatics | 0.033 | 3.732 | 0.107 | 0.405 |
| Hedges | -0.321 | -0.455 | -0.129 | -0.081 |
| Negative affect | -1.980 | -6.889 | **-1.235** | **-0.827** |
| Con polarity | -2.723 | -17.095 | **-1.730** | **-2.011** |

Figure 9 shows how both models rely on the studied linguistic features similarly, allowing us to validate trough this dual-analysis how linguistic features impact on stance classification, although also showing how both models differ. We can then conclude that pro polarity, positive affect, con polarity, and negative affect are among the most influential in both models.

**Figure 9**

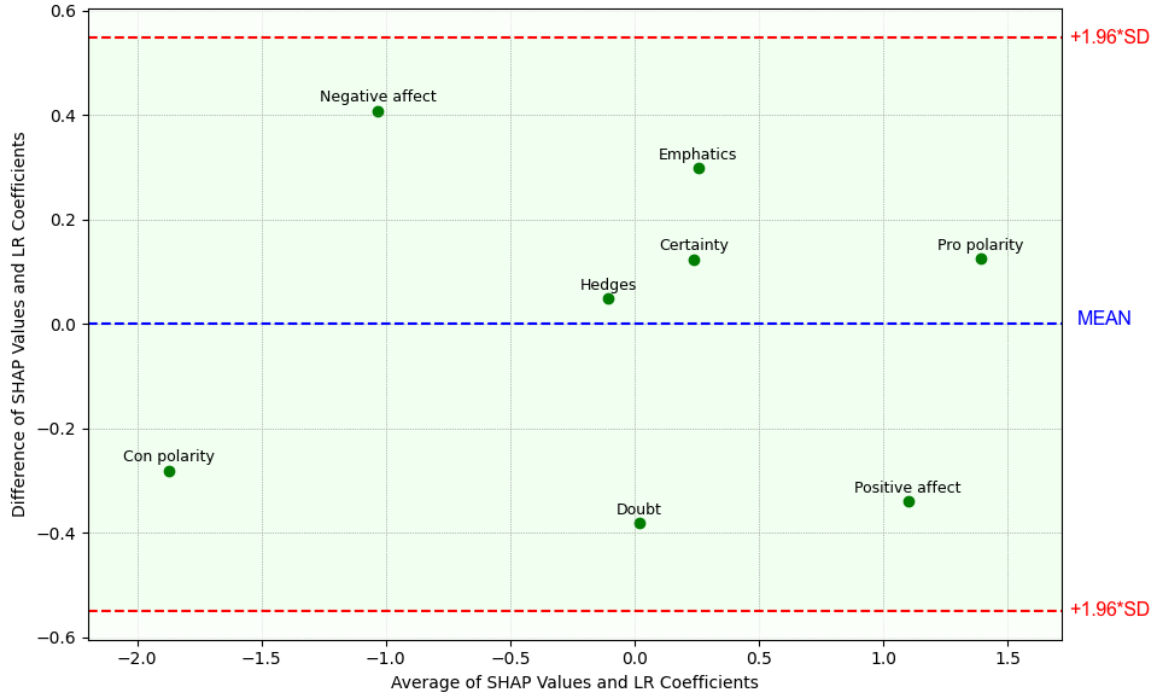
*Comparison of normalized SHAP values and Logistic Regression Coefficients.*




A two-sample Welch's t-test was conducted to compare the means of LR $\beta$ and SHAP values. The results indicated that there was no significant difference between the means of the two groups, t (14.00) = $2.08 \times 10^{-16}$, p = 1.0. The Bland-Altman plot in Figure 10 shows StanceFit (SHAP) values and StanceFeat (LR) coefficients for each feature, indicating the relative importance or magnitude of a feature assessed by both methods. The vertical axis represents the difference between SHAP values and LR coefficients, showing the discrepancy in the evaluation of each feature between both methods. Features above the MEAN line (negative affect, emphatics, certainty, hedges, pro polarity) indicate that SHAP assigns a higher value than LR.

**Figure 10**

*Bland-Altman Plot of Comparison of SHAP values and Logistic Regression Coefficients.*




Opposely, features below the MEAN line (con polarity, doubt, positive affect) indicate that LR assigns a higher value than SHAP. The dashed lines represent the limits of agreement, set at ±1.96 *SD* (95%) from the mean difference, indicating the range within which most differences between the two methods should lie if they are considered in reasonable agreement. The fact that hedges is positioned very close to the MEAN line suggests that both models agree closely on assessing this feature with minimal bias. Features positioned towards the left side of the plot (con polarity, negative affect) indicate a lower average importance assessed by both models, while features towards the right (pro polarity, positive affect) suggest a higher average importance.

# 5. Conclusions

The StanceSentences dataset is a ground truth for any dataset on subjective stance at the sentence level and could be the seed to capture more complex similar expressions. Seeing the trend of few-shot learning models in NLP, we consider that StanceSentences sets a benchmark in the field, particularly concerning the dataset size and stratification (1,280 examples perfectly class balanced), which is crucial for effective stance classification study. Answering RQ1, the study of SetFit using different lenses to observe its adaptability in the subjective stance classification task confirms that few-shot learning is highly effective, which is visually observable through the t-SNE plots. Its variation paraphrase-mpnet-base-v2 seems to have contributed to understanding more nuanced expressions of subjective stance (seen in Example 2, where the verb "*deter*" was used to define the oppose classification even when that word was present two times in the dataset).

The choice of altering the feature aggregation proposed by Biber and Finegan (1989), in which we disaggregated emphatics and hedges from certainty and doubt features, respectively, added a feature dimensionality ad-hoc for the analysis of the political language. The pertinance of the decision is supported by the results of the SHAP analysis, where doubt and hedges became predictors of the oppose class (Table 10), contrary to certainty and emphatics, which resulted in better predictors for the support class. As mentioned before, during the feature engineering stage, we found enough terms indicative that hedges and

emphatics could be disaggregated safety. This claim is observable in Figure 4, in which certainty and emphatics high an enormously disproportionate number of occurrences compared to doubt and hedges, meaning that, although numerous in quantity, the neural model did not use certainty and emphatics as important predictors, given their low SHAP values. Therefore, answering RQ2, we can conclude that while there are some biases and variances in how both models assess the features, there is an overall agreement that the most important linguistic features to predict the classification of support or oppose stance are pro polarity, con polarity, positive affect, and negative affect; and that emphatics and doubt may be used as complementary features toward support and oppose respectively. The importance of the affective dimension of stance is observable alongside this study, and although many previous researches support this finding, we present quantitative data that may guide future research on subjective stance.

This study also proves the importance of incorporating traditional linguistic methods, such as corpus linguistics, to bring explainability to LLMs. In this sense, although SHAP behaved accurately, giving explanations at the token level, further studies should focus on the observed behavior of SetFit clustering tokens into phrases (present in Examples 1 and 2) since subtlety and implicitness are usually conveyed by articulating tokens in subjective stance expressions.

Finally, this research brings limitations: (1) The opinionated decisions made in the classification of terms to build the lexicons could be debatable since many decisions were made in the context of the examples in the StanceSentences dataset, therefore the lexicons proposed should be analyzed and improved in a broader context; (2) StanceFit needs to be benchmarked in its generalization capabilities with data from other domains and with more complex expressions of subjective stance; (3) although our focus on specific POS is appropriate for our research goals, this research omitted the analysis of important rhetorical resources—used for opposition in its intention to soften confrontation—for instance, the preposition "*against*" and nouns like ("*aggression*", "*fight*", "*struggle*", Etc.); therefore the study of other POS could be expanded in further research; and (4) although in this research we operationalized features most of the time using lexicogrammatical artifacts and methods, (subjective) stance is complicated, and we acknowledge that it needs a multi-layered analysis, at the pragmatics and semantics levels; for instance the topic of "*war*" and the militaristic language in the oppose class needs a more nuanced analysis of how they interact with each other.

# Appendix

List of features and terms investigated using a multidimensional approach combining (1) lexicons and (2) spaCy's linguistic features, combining morphosyntactic capabilities and pattern matchers. All the lexicon entries are expressed in their lemma form and without contractions; for instance, the phrase "*go to*" is the lemma of "*going to*" and "*can not*" is the lemma of "*can't*" or "*cannot*".

## Affect

1. Positive adjectives

   *advanced, affordable*, *agile*, *ambitious*, *antitrust*, *assistive*, *attractive*, *available*, *balanced*, *bilateral*, *bipartisan*, *bold*, *capable*, *caregiving*, *clean*, *collective*, *competitive*,

*comprehensive*, *constructive*, *cooperative*, *dedicated*, *democratic*, *diplomatic*, *durable*, *effective*, *efficient*, *eligible*, *equitable*, *essential*, *excellent*, *excited*, *extraordinary*, *fair*, *fast*, *fellow*, *fortunate*, *founding*, *free*, *friendly*, *functional*, *good*, *happy*, *hardworking*, *honest*, *honorable*, *hopeful*, *hospitable*, *humanitarian*, *important*, *inclusive*, *incredible*, *independent*, *indispensable*, *innocent*, *innovative*, *intellectual*, *interconnected*, *interested*, *kind*, *legal*, *legitimate*, *live*, *loved*, *magnificent*, *meaningful*, *moral*, *multilateral*, *multinational*, *mutual*, *necessary*, *nimble*, *peaceful*, *pleased*, *popular*, *positive*, *powerful*, *practical*, *prepared*, *pretty*, *productive*, *profitable*, *proper*, *prosperous*, *proud*, *qualified*, *remarkable*, *resilient*, *responsible*, *right*, *rightful*, *safe*, *satisfied*, *sincere*, *singular*, *smart*, *sovereign*, *special*, *stable*, *steadfast*, *stimulant*, *strategic*, *strong*, *substantial*, *successful*, *super*, *tolerated*, *transformative*, *transparent*, *unashamed*, *uncensored*, *vibrant*, *vital*, *well*, *willing*, *wonderful*, *young*

*top-of-the-line*, *up-to-date*

*non-*/*not* [negative adjective]

2. Positive adverbs

   *better*, *commercially*, *democratically*, *effectively*, *enthusiastically*, *environmentally*, *fairly*, *fortunately*, *freely*, *good*, *judiciously*, *mutually*, *responsibly*, *rightly*, *successfully*, *tirelessly*, *traditionally*, *understandably*, *wisely*

   *not* [negative adverb]

3. Positive verbs

   *accomplish*, *achieve*, *address*, *alleviate*, *aspire*, *build*, *care*, *cherish*, *clean*, *clear*, *coordinate*, *count*, *create*, *cure*, *decriminalize*, *discover*, *drive*, *educate*, *empower*, *encourage*, *endure*, *engage*, *enhance*, *enjoy*, *enrich*, *envision*, *excel*, *fix*, *flourish*, *forge*, *fulfil*, *guide*, *hearten*, *immunize*, *implement*, *improve*, *install*, *like*, *love*, *materialize*, *modernize*, *organize*, *please*, *plow*, *preserve*, *produce*, *promise*, *prosper*, *qualify*, *rally*, *reach*, *rebuild*, *recover*, *reform*, *refresh*, *relieve*, *relish*, *renew*, *repair*, *rescue*, *resolve*, *restore*, *revamp*, *revise*, *revitalize*, *reward*, *rid*, *satisfy*, *save*, *secure*, *serve*, *soften*, *stimulate*, *streamline*, *strengthen*, *strive*, *succeed*, *suit*, *surprise*, *thank*, *thrive*, *tolerate*, *train*, *transform*, *upgrade*, *win*, *wish*, *work*

   *not* [negative verb]

4. Negative adjectives

   *abhorrent*, *affected*, *afraid*, *aggressive*, *alarmed*, *alone*, *angry*, *anti*, *anxious*, *arduous*, *ashamed*, *authoritarian*, *bad*, *brutal*, *burdensome*, *catastrophic*, *colonial*, *complex*, *contagious*, *contaminated*, *controversial*, *criminal*, *cruel*, *dangerous*, *dark*, *deadly*, *deliberate*, *dependent*, *despicable*, *deteriorated*, *devastating*, *difficult*, *dire*, *disastrous*, *discriminatory*, *disturbed*, *divorced*, *embarrassed*, *enslaved*, *evil*, *exhausted*, *extremist*, *faulty*, *flagrant*, *frightened*, *frightening*, *grave*, *grim*, *guilty*, *harmful*, *harsh*, *hateful*, *horrible*, *ill*, *illegal*, *illicit*, *immoral*, *imperialistic*, *indifferent*, *infectious*, *isolated*, *jeopardized*, *killer*, *lethal*, *malicious*, *malign*, *misguided*, *negative*, *numb*, *odd*, *overdue*, *phony*, *poor*, *precarious*, *racist*, *rampant*, *rash*, *sectarian*, *selfish*, *severe*, *systemic*, *terrible*, *terrorist*, *threatening*, *tough*, *toxic*, *tragic*, *troublesome*, *turbulent*, *unauthorized*, *uncivil*, *unconstitutional*, *uncontrolled*, *underserved*, *unending*, *unfair*, *unfortunate*, *unilateral*, *unjustified*, *unnecessary*, *unpopular*, *unprovoked*, *unrealistic*,

*unrelenting*, *unsafe*, *unwise*, *urgent*, *vicious*, *violent*, *vulnerable*, *weak*, *worried*, *worse*, *worst*, *wrong*

*at risk*, *in jeopardy*, *under attack*, *under siege*, *under threat*

*not* [positive adjective]

5. Negative adverbs

   *aggressively*, *arbitrarily*, *badly*, *disturbingly*, *negatively*, *overwhelmingly*, *painstakingly*, *sadly*, *unfortunately*, *unjustly*

   *not* [positive adverb].

6. Negative verbs

   *abandon*, *aggravate*, *bear*, *break*, *burn*, *bury*, *cause*, *censor*, *cheat*, *compromise*, *concern*, *confuse*, *cope*, *criminalize*, *crumble*, *damage*, *delay*, *demand*, *depend*, *destabilize*, *destroy*, *detain*, *deteriorate*, *dethrone*, *die*, *discourage*, *dismantle*, *distract*, *disturb*, *divide*, *dump*, *endanger*, *enslave*, *expose*, *fail*, *fear*, *forget*, *hang*, *harm*, *hate*, *hurt*, *imperil*, *impose*, *inconvenience*, *intimidate*, *invade*, *jeopardize*, *kill*, *loom*, *lose*, *mislead*, *misrepresent*, *misuse*, *nuclearize*, *overthrow*, *overturn*, *overwhelm*, *pain*, *pay*, *pose*, *possess*, *precipitate*, *prey*, *rage*, *resent*, *reshore*, *sacrifice*, *scourge*, *spend*, *spiral*, *steal*, *stir*, *stricken*, *stumble*, *suffer*, *tear*, *traffic*, *trouble*, *violate*, *violent*, *war*, *waste*, *worry*, *wrack*, *wrench*

   *not* [positive verb]

## Evidentiality

7. Certainty adjectives

   *absolute*, *accountable*, *affirmative*, *attentive*, *aware*, *certain*, *cognizant*, *complete*, *concerted*, *concrete*, *conducive*, *confident*, *continued*, *convinced*, *credible*, *determined*, *direct*, *distinct*, *entire*, *evident*, *explicit*, *feasible*, *final*, *firm*, *flat*, *frank*, *guaranteed*, *inconceivable*, *inevitable*, *infallible*, *inherent*, *institutional*, *known*, *material*, *objective*, *obvious*, *only*, *particular*, *patent*, *precise*, *ready*, *real*, *realistic*, *reliable*, *resolute*, *resolved*, *same*, *secure*, *structural*, *substantive*, *sure*, *sustainable*, *tangible*, *true*, *unambiguous*, *unanimous*, *unarguable*, *unavoidable*, *unchanging*, *unconditional*, *whole*

   *well-known*

   *not* [doubt adjective]

8. Certainty adverbs

   *absolutely*, *actually*, *all*, *already*, *always*, *anymore*, *anywhere*, *certainly*, *clearly*, *completely*, *definitely*, *entirely*, *especially*, *ever*, *everywhere*, *exactly*, *explicitly*, *finally*, *firmly*, *forever*, *frankly*, *fully*, *hence*, *here*, *immediately*, *indeed*, *just*, *never*, *obviously*, *often*, *once*, *particularly*, *personally*, *precisely*, *quickly*, *really*, *seriously*, *strictly*, *surely*, *systemically*, *then*, *therefore*, *thus*, *too*, *totally*, *truly*, *twice*, *ultimately*, *up*, *well*

   *of course*, *without a doubt*

   *not* [doubt adverb]

9. Certainty verbs

*accept*, *acknowledge*, *affirm*, *aim*, *announce*, *assure*, *attest*, *believe*, *conclude*, *confirm*, *convince*, *define*, *demonstrate*, *designate*, *determine*, *ensure*, *establish*, *explain*, *express*, *illustrate*, *include*, *inform*, *introduce*, *know*, *learn*, *note*, *notify*, *order*, *outline*, *perceive*, *present*, *prove*, *reaffirm*, *realize*, *recognise*, *recognize*, *rely*, *show*, *solidify*, *state*, *swear*, *understand*, *uphold*

*make sure*

*not* [doubt verb], *has*/*have shown*

10. Doubt adjectives

    *ambiguous*, *confused*, *covert*, *distrustful*, *hypothetical*, *impossible*, *possible*, *remote*, *uncertain*, *undetermined*, *untrue*, *usual*

    *not* [certainty adjectives]

11. Doubt adverbs

    *eventually*, *perhaps*, *possibly*, *probably*

    *not* [certainty adverb]

12. Doubt verbs

    *appear*, *attempt*, *challenge*, *distrust*, *expect*, *feel*, *hope*, *imply*, *indicate*, *intend*, *prefer*, *question*, *seem*, *sense*, *think*, *try*, *waver*

    *not* [certainty verb]

13. Emphatics adjectives

    *active*, *any*, *basic*, *big*, *bottom*, *broad*, *central*, *clear*, *close*, *common*, *consistent*, *countless*, *critical*, *crucial*, *current*, *decisive*, *deep*, *detailed*, *disadvantaged*, *dominant*, *dramatic*, *early*, *easy*, *economical*, *elementary*, *emotional*, *endless*, *enormous*, *enough*, *equal*, *eternal*, *even*, *fantastic*, *first*, *focused*, *foremost*, *former*, *front*, *full*, *fundamental*, *further*, *general*, *global*, *great*, *groundbreaking*, *hard*, *heavy*, *high*, *historic*, *huge*, *immediate*, *individual*, *individualized*, *instant*, *integral*, *integrated*, *intense*, *just*, *key*, *landmark*, *large*, *last*, *lasting*, *leading*, *legendary*, *light*, *long*, *longstanding*, *major*, *many*, *mass*, *massive*, *maximum*, *middle*, *minimum*, *modern*, *more*, *most*, *much*, *multiple*, *multiplier*, *narrow*, *new*, *next*, *old*, *ongoing*, *open*, *other*, *outer*, *overall*, *own*, *parallel*, *past*, *persistent*, *pivotal*, *present*, *previous*, *primary*, *prominent*, *prompt*, *pursuant*, *quick*, *recent*, *reliant*, *rich*, *rising*, *seamless*, *serious*, *shared*, *short*, *significant*, *similar*, *simple*, *single*, *small*, *solid*, *sophisticated*, *spare*, *specific*, *steady*, *such*, *sweeping*, *targeted*, *tectonic*, *top*, *total*, *tremendous*, *ultimate*, *uniform*, *universal*, *unparalleled*, *unprecedented*, *upcoming*, *very*, *vigorous*, *visible*, *wide*, *worldwide*, *worthwhile*

    *long-range*

    *not* [hedges adjective]

    *real*/*so* [adjective]

14. Emphatics adverbs

    *actively*, *again*, *ahead*, *alone*, *also*, *anew*, *away*, *before*, *broadly*, *currently*, *deeply*, *directly*, *down*, *early*, *enough*, *equally*, *even*, *exceptionally*, *extensively*, *extremely*, *far*, *fast*, *foremost*, *forth*, *forward*, *fundamentally*, *further*, *furthermore*, *globally*, *hard*, *hardly*, *highly*, *historically*, *immeasurably*, *importantly*, *increasingly*, *incredibly*, *independently*, *instinctively*, *largely*, *long*, *longer*, *more*, *most*, *much*, *nevertheless*, *now*,

*only*, *outright*, *over*, *overall*, *overly*, *philosophically*, *politically*, *presently*, *primarily*, *principally*, *privately*, *profoundly*, *quick*, *quietly*, *quite*, *rapidly*, *recently*, *repeatedly*, *right*, *seamlessly*, *severely*, *shortly*, *significantly*, *simply*, *simultaneously*, *slowly*, *so*, *soon*, *still*, *strongly*, *swiftly*, *very*, *vigorously*, *within*, *worldwide*, *yet*

*a lot*, *at the top*, *for sure*, *in addition*, *in fact*, *in particular*, *in reality*

*not* [hedges verb]

*as I/we* (*have*) ["*say*", "*note*", "*announce*"]

*so* [adjective]

15. Emphatics verbs

    *accelerate*, *acquire*, *amplify*, *anticipate*, *arise*, *articulate*, *augment*, *become*, *begin*, *boost*, *bridge*, *bring*, *broaden*, *broker*, *center*, *close*, *complete*, *concentrate*, *condition*, *connect*, *consolidate*, *consume*, *continue*, *converge*, *declare*, *deepen*, *disrupt*, *double*, *earn*, *elevate*, *emerge*, *emphasize*, *enforce*, *enter*, *escalate*, *evolve*, *exaggerate*, *exceed*, *excite*, *exercise*, *expand*, *exploit*, *extend*, *focus*, *force*, *generate*, *ground*, *grow*, *harden*, *highlight*, *identify*, *impact*, *increase*, *insist*, *integrate*, *intensify*, *jump*, *keep*, *lead*, *leverage*, *lift*, *maintain*, *mobilize*, *multiply*, *open*, *orient*, *plead*, *point*, *prepare*, *prioritise*, *prioritize*, *pursue*, *push*, *raise*, *rededicate*, *redefine*, *redouble*, *reignite*, *repeat*, *rise*, *spur*, *take*, *target*, *transition*, *undertake*, *unleash*, *value*

    *call for*, *fan the flames*, *make clear*, *point out*

    *not* [hedges verb]

    *do* [verb]

16. Hedges adjectives

    *additional*, *alternative*, *appropriate*, *convenient*, *down*, *few*, *less*, *likely*, *little*, *low*, *moderate*, *nuanced*, *potential*, *preventative*, *reasonable*, *reduced*, *relevant*, *satisfactory*, *several*, *some*, *various*

    *not* [emphatics adjective]

17. Hedges adverbs

    *about*, *almost*, *alternatively*, *anyway*, *around*, *but*, *closely*, *elsewhere*, *except*, *generally*, *however*, *indirectly*, *initially*, *instead*, *later*, *like*, *little*, *maybe*, *mostly*, *nearly*, *partly*, *potentially*, *pretty*, *rather*, *regardless*, *somewhat*, *though*, *virtually*

    *a bit*, *a little*, *in a way*, *in general*, *in part*, *in principle*, *kind of*, *pretty much*, *sort of*

    *not* [emphatics adverb]

18. Hedges verbs

    *avoid*, *creep*, *cut*, *decline*, *degrade*, *deter*, *dilute*, *diminish*, *freeze*, *grind*, *hide*, *hinder*, *impair*, *interrupt*, *isolate*, *lessen*, *lower*, *near*, *reduce*, *restrict*, *rot*, *suppress*, *undermine*, *underscore*, *weaken*

    *hold down*, *might be*

## Polarity

19. Pro adjectives

*accepted*, *agreed*, *allowed*, *backed*, *bound*, *brought*, *built*, *commended*, *committed*, *contributed*, *coordinated*, *created*, *defended*, *defensive*, *developed*, *enabled*, *encouraged*, *enriched*, *equipped*, *expanded*, *favorable*, *funded*, *grown*, *helpful*, *implemented*, *improved*, *included*, *invested*, *invited*, *joined*, *joint*, *maintained*, *offered*, *pledged*, *prioritized*, *provided*, *raised*, *reaffirmed*, *recognized*, *renewed*, *respected*, *sponsored*, *stood*, *strengthened*, *supportive*, *sustained*, *unleashed*, *welcomed*

*not* [con adjective]

[adjective] [*commitment*, *support*, *help*, *endorsement*, *favor*, *agreement*]

20. Pro adverbs

    *together*

    *on board*

    *not* [con adverb]

21. Pro verbs

    *accord*, *advance*, *advocate*, *agree*, *aid*, *allow*, *ally*, *answer*, *applaud*, *approve*, *assist*, *associate*, *back*, *benefit*, *bind*, *bolster*, *carry*, *champion*, *coddle*, *cohost*, *collaborate*, *combine*, *commend*, *commit*, *contribute*, *cooperate*, *cultivate*, *defend*, *develop*, *embrace*, *enable*, *endorse*, *equip*, *facilitate*, *favor*, *finance*, *find*, *foster*, *fund*, *further*, *give*, *guarantee*, *guard*, *help*, *incentivise*, *insure*, *invest*, *invite*, *join*, *negotiate*, *nurture*, *offer*, *partner*, *permit*, *pledge*, *position*, *praise*, *progress*, *promote*, *propose*, *protect*, *provide*, *pump*, *recommend*, *reinforce*, *rekindle*, *respect*, *respond*, *safeguard*, *seek*, *sponsor*, *stand*, *supply*, *support*, *sustain*, *trust*, *unite*, *welcome*

    *believe in*, *look forward*, *pursuit of*, *stand for*

    *not* [con verb]

    *make* (*a*) *commitment*

    *have* ([*a*, *an*, *the*]) *interest*

    *recognize* (*the*) *importance*

    *fight for* — where *fight* is a verb

22. Con adjectives

    *accused*, *compromised*, *concerned*, *concerning*, *condemned*, *confronted*, *contrary, counter*, *criticized*, *cut*, *defeated*, *denounced*, *deplored*, *deterred*, *diminished*, *divided*, *ended*, *exploited*, *exposed*, *faced*, *forced*, *fought*, *hurt*, *imposed*, *interrupted*, *lessened*, *limited*, *misled*, *opposed*, *posed*, *prevented*, *refuted*, *suffered*, *suppressed*, *threatened*, *unacceptable*, *unwilling*, *vigilant*, *weakened*, *wracked*

    *angry at*, *broken down*, *broken-down*, *held down*, *held-down*

    *not* [pro adjective]

    [adjective] [*concern*, *opposition*, *threat*]

    [*a matter of*, *more than a*, *a list of*] *concern*

    [adjective] *against*

23. Con adverbs

    *back*, *detrimentally*, *no*, *not*, *off*, *out*, *outside*

*in opposition*

*not* [pro adverb]

[adverb] *against*

[adverb] [*concern*, *opposition*, *disagreement*, *thread*]

24. Con verbs

*accuse*, *affect*, *argue*, *arrest*, *attack*, *battle*, *beat*, *blame*, *buck*, *clash*, *combat*, *condemn*, *confront*, *contradict*, *control*, *counter*, *curtail*, *deal*, *dedicate*, *defeat*, *defy*, *denounce*, *deplore*, *destruct*, *devote*, *disagree*, *disallow*, *disapprove*, *discontinue*, *dispute*, *dissent*, *eliminate*, *end*, *exclude*, *face*, *fight*, *finish*, *haunt*, *interfere*, *limit*, *muzzle*, *object*, *oppose*, *outlaw*, *pit*, *press*, *prevent*, *prosecute*, *punish*, *refuse*, *refute*, *regulate*, *reject*, *resist*, *shut*, *stop*, *strike*, *struggle*, *suspend*, *tackle*, *threaten*, *wage*

*break down*, *can not*, *spark concern*, *take on*

*not* [pro verb]

[verb] ([*a*, *an*, *the*, *more of a*]) [*concern*, *opposition*, *disagreement*, *thread*]

[verb] *against*

### Modality

25. Predictive modal

*shall*, *will*, *would*

*go to*

26. Possibility modal

*can*, *could*, *may*, *might*

27. Necessity modal

*must*, *should*

*have to*, *need to*, *ought to*

## References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD*. https://doi.org/10.48550/arXiv.1907.10902.
2. AlDayel, A., & Magdy, W. (2021). Stance Detection on Social Media: State of the Art and Trends. *arXiv*. https://doi.org/10.48550/arXiv.2006.03644.
3. Allaway, E., & McKeown, K. (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)*, 8913-8931. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.717.
4. Alturayeif, N., Luqman, H. & Ahmed, M. (2023). A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing & Applications*, 35, 5113-5144. https://doi.org/10.1007/s00521-023-08285-7.

5. Anand, P., Walker, M. A., Abbott, R., Fox Tree, J. E., Bowmani, R., & Minor, M. (2011). Cats Rule and Dogs Drool!: Classifying Stance in Online Debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, 1-9. Association for Computational Linguistics. https://aclanthology.org/W11-1701.
6. Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse*, 9(1), 93-124. https://doi.org/10.1515/text.1.1989.9.1.931.
7. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. Longman.
8. Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.
9. Chauhan, D., Kumar, R., & Ekbal, A. (2019). Attention Based Shared Representation for Multi-task Stance Detection and Sentiment Analysis. In *International Conference on Neural Information Processing (ICONIP)*, 661-669. Springer, Cham. https://doi.org/10.1007/978-3-030-36802-9_70.
10. Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., & Roth, D. (2019). Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*, 1, 542-557. https://doi.org/10.48550/arXiv.1906.03538.
11. Chilton, P. (2004). *Analyzing Political Discourse: Theory and Practice*. Routledge.
12. Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F., & Collier, N. (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*, 1715-1724. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.157.
13. Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F., & Collier, N. (2022). Incorporating Stock Market Signals for Twitter Stance Setection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4074-4091. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.281.
14. Darwish, K., Magdy, W., & Zanouda, T. (2017). Trump vs. Hillary: What went Viral during the 2016 US Presidential Election. *arXiv*. https://doi.org/10.48550/arXiv.1707.03375.
15. Du, J., Xu, R., He, Y., & Gui, L. (2017). Stance Classification with Target-specific Neural Attention Networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3988-3994. https://doi.org/10.24963/ijcai.2017/557.
16. Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction* (pp. 139-182). John Benjamins Publishing Company.

17. Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. 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)*, 3816-3830. https://doi.org/10.18653/v1/2021.acl-long.295.

18. Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. 1859-1874. Association for Computational Linguistics.

19. Hasan, K. S., & Ng, V. (2013). Frame Semantics for Stance Classification. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning* (pp. 124-132). Association for Computational Linguistics. https://aclanthology.org/W13-3514.

20. Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

21. Jayaram, V., & Allaway, E. (2021). Human Rationales as Attribution Priors for Explainable Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5540-5554. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.450.

22. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/.

23. Kawintiranon, K., & Singh, L. (2021). Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4725-4735. Association for Computational Linguistics. https://doi.org/10.18653/V1/2021.NAACL-MAIN.376.

24. Khamkhien, A. (2014). LINGUISTIC FEATURES OF EVALUATIVE STANCE: FINDINGS FROM RESEARCH ARTICLE DISCUSSIONS. *Indonesian Journal of Applied Linguistics*, 4(1), 54-69. https://doi.org/10.17509/ijal.v4i1.600.

25. Lai, M., Cignarella, A., Hernandez, D., Bosco, C., Patti, V., & Rosso, P. (2020a). Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63, 101075. https://doi.org/10.1016/j.csl.2020.101075.

26. Lai, M., Patti, V., Ruffo, G., & Rosso, P. (2020b). Brexit: Leave or remain? The role of user's community and diachronic evolution on stance detection. *Journal of Intelligent & Fuzzy System*, 39(2), 2341-2352. https://doi.org/10.3233/JIFS-179895.

27. Li, Y., & Caragea, C. (2019). Multi-task Stance Detection with Sentiment and Stance Lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6299-6305. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1657.

28. Li, Y., Sosea, T., Sawant, A., Nair, A. J., Inkpen, D., & Caragea, C. (2021). P-stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2355-2365). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.208.

29. Li, Y., Wang, S., Lin, C., Guerin, F., & Barrault, L. (2023). FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1558-1563). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.114.

30. Martin, J. R., & White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

31. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3945-3952. European Language Resources Association (ELRA).

32. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Marquez, L., & Moschitti, A. (2018). Automatic Stance Detection Using End-to-End Memory Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 767-776. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1070.

33. Nielsen, F. A. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages* (pp. 93-98). https://doi.org/10.48550/arXiv.1103.2903.

34. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019). Explanation of Machine Learning Models Using Improved Shapley Additive Explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 546. https://doi.org/10.1145/3307339.3343255.

35. OpenAI. (2024). ChatGPT [Computer software]. https://www.openai.com.

36. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

37. Peters, G., & Woolley, J. T. (n.d.). The American Presidency Project. University of California, Santa Barbara. https://www.presidency.ucsb.edu/.

38. Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2019). STANCY: Stance Classification Based on Consistency Cues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6413-6418. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1675.

39. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. *arXiv*. https://arxiv.org/abs/1908.10084.

40. Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.

41. Saha, R. R., Lakshmanan, L. V. S., & Ng, R. (2023). Stance Detection with Explanations. *Computational Linguistics*, 50(1), 193-235. https://doi.org/10.1162/coli_a_00501.

42. Sobhani, P., Mohammad, S., & Kiritchenko, S. (2016). Detecting Stance in Tweets and Analyzing its Interaction with Sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 159-169). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-2021.

43. Sobhani, P., Inkpen, D., & Zhu, X. (2017). A Dataset for Multi-Target Stance Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 551-557). Association for Computational Linguistics.

44. Somasundaran, S., & Wiebe, J. (2010). Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 116-124). Association for Computational Linguistics.

45. Sun, Q., Wang, Z., Zhu, Q., & Zhou, G. (2016). Exploring Various Linguistic Features for Stance Detection. *Natural Language Understanding and Intelligent Applications. ICCPOL 2016, NLPCC 2016*, 840-847. Springer, Cham. https://doi.org/10.1007/978-3-319-50496-4_76.

46. Sun, Q., Wang, Z., Li, S., Zhu, Q., & Zhou, G. (2019). Stance detection via sentiment information and neural network model. *Frontiers of Computer Science*, 13(1), 127-138. https://doi.org/10.1007/s11704-018-7150-9.

47. Sun, Q., Wang, Z., Zhu, Q., & Zhou, G. (2018). Stance Detection with Hierarchical Attention Network. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2399-2409). Association for Computational Linguistics.

48. The pandas development team (2020). pandas-dev/pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134.

49. Tunstall, L., Reimers, N., Seo Jo, U. E., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. *arXiv*. https://arxiv.org/abs/2209.11055.

50. Vychegzhanin, S., & Kotelnikov, E. (2021). A New Method for Stance Detection Based on Feature Selection Techniques and Ensembles of Classifiers. *IEEE Access*, 9, 134899-134915. https://doi.org/10.1109/ACCESS.2021.3116657.

51. Walker, M., Anand, P., Abbott, R., Tree, J. E., Martell, C., & King, J. (2012). That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53, 719-729.

52. Wang, Z., Sun, Q., Li, S., Zhu, Q., & Zhou, G. (2020). Neural Stance Detection With Hierarchical Linguistic Representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 635-645. https://doi.org/10.1109/TASLP.2020.2963954.

53. Wasserblat, M. (2021, December 14). Sentence Transformer Fine-Tuning (SetFit). Intel Community. https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Sentence-Transformer-Fine-Tuning-SetFit/post/1407712.

54. Wilson, J. (1990). *Politically Speaking: The Pragmatic Analysis of Political Language*. Basil Blackwell.

55. Zhang, Y., Ding, D., & Jing, L. (2022). How would stance detection techniques evolve after the launch of ChatGPT? *arXiv*. https://arxiv.org/abs/2212.14548.