Juan-Francisco Reyes

Explainable Subjective Stance Classification with SetFit in Political Discourse

Stance classification in *Natural Language Processing* (*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 sourced from professional politicians for binary subjective stance classification – *support* or *oppose* – using *bootstrapping* in a *SetFit* model. 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 *positive affect*, *negative affect*, *pro polarity*, *con polarity*, *certainty*, *emphatics*, *doubt*, *hedges*. *SHAP* analysis quantifies the influence of these features on SetFit model decisions. Our findings demonstrate that iterative bootstrapping significantly enhances the efficacy of few-shot learning in subjective stance classification, and we 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.

1. Introduction

Stance classification in NLP helps identify support or opposition in political discourse. This study focuses on identifying linguistic markers that predict subjective stance – defined as expressions of support or opposition – toward political targets or policy issues. This approach establishes a foundation for understanding the linguistics of stance and evaluating the methodology's effectiveness. While stance classification has advanced significantly, current research often exhibits a disconnection between the granularity of available data and the precise linguistic mechanics of political discourse. Existing approaches frequently operate at broader, multi-sentence levels of analysis, making it difficult to see the lexicogrammatical ways in which stance is actually constructed.

To address these challenges, our research explicitly isolates subjective stance at the sentence level to provide a controlled linguistic environment. Specifically, we construct a sentence-level dataset of speech from professional politicians for binary stance classification – *support* or *oppose* – using a bootstrapping method within the few-shot learning framework of *SetFit* (Tunstall et al., 2022), iteratively refining the model. This approach aligns with the difficulty of collecting and annotating political discourse sentences and tests whether SetFit can analyze complex political language with limited data. SetFit, with its advanced use of the *Sentence Transformers* architecture

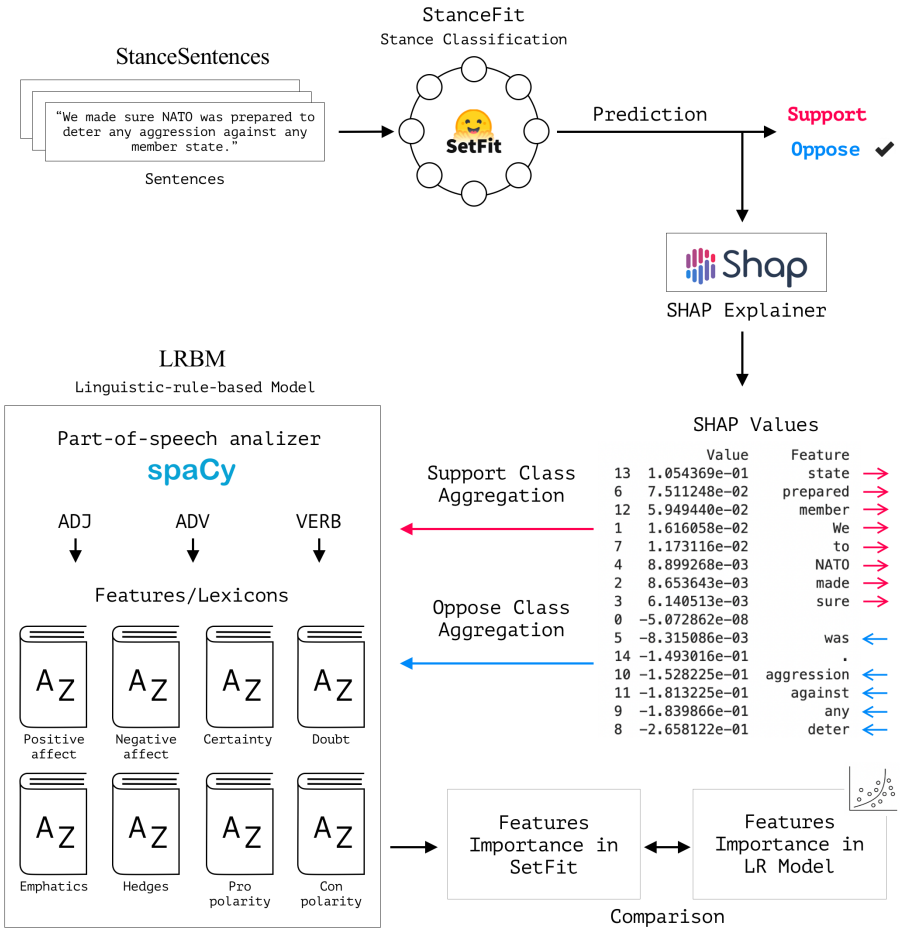

Figure 1: NLP Explainability Process Using SHAP on StanceFit, the SetFit model.

(Reimers & Gurevych, 2019), should be proficient at handling these subtleties, decoding everything from explicit statements to the more nuanced shades of expression, ensuring a thorough and comprehensive analysis.

To address concerns about the explainability of neural models, this study incorporates traditional linguistic approaches – corpus linguistics methods, tailored lexicons, and lexicogrammatical rules – to decode how models use specific word choices for stance classification, focusing on eight linguistic features: *positive affect*, *negative affect*, *pro polarity*, *con polarity*, *certainty*, *emphatics*,

doubt, and *hedges*. 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 et al., 2020), 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. Figure 1 provides a visual overview of this methodological pipeline, illustrating the progression from the initial dataset bootstrapping to linguistic feature extraction and the subsequent parallel classification and explainability approaches.

This study aims to answer questions about stance classification in political discourse:

- **RQ1:** How does the SetFit (few-shot) model evolve during the bootstrapping process for building a stratified, explainable dataset for sentence-level stance classification in political discourse?
- **RQ2:** What linguistic features are most important in predicting the classification of *support* or *oppose* stance?

The contribution of this work is three-fold: (1) the introduction of a dataset for few-shot learning models for classifying binary stance expressions from professional politicians 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 treat stance classification and stance detection similarly, 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

Table 1: Taxonomy of **S**tance **F**eatures 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 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.

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 acknowledging 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 how 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 (Table 1) aids in understanding the multifaceted nature of stance and provides a comprehensive approach to analyzing it in text.

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 and Magdy (2021), Chauhan et al. (2019), Lai, Cignarella, et al. (2020), Mohammad et al. (2016), Sobhani et al. (2016), Somasundaran and 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 argument-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.

Similarly, 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. Also, Sun et al. (2016) explored four linguistic features, including lexical, morphological, semantic, and syntactic 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, Cignarella, et al. (2020) 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.

P. Chen et al. (2021) used a BERT model for stance classification complemented with explicit n-gram features and handcrafted or linguistic features (like n-grams, part-of-speech cues, and subjectivity cues) for stance detection, improving interpretability. Gómez-Suta et al. (2023) used Latent Dirichlet Allocation (LDA) and traditional NLP techniques to improve and explain stance detection, relying on human-interpretable topic features from tweets.

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 (LLMs), 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 linguistic 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.

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 (S. 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), determining how the relationships between different phrases influence stance detection.

Further refining the granularity of the explanation, Jayaram and Allaway (2021) applied an *MAW* (*Mean Attention Weights*) method to the *VAST* dataset (Allaway & McKeown, 2020), focusing on the token level. This method delves into the individual word or token, offering precise insights into how specific words contribute to the detected stance.

More recent studies explicitly focus on generating passage-level explanations. Y. Zhang et al. (2022) evaluated *ChatGPT*'s ability to provide natural language rationales for the *SemEval-2016* and *P-Stance* datasets. B. Zhang et al. (2023) used rule-entity binding on an X Posts dataset to provide explicit rules as explanations. Saha et al. (2024) generated human-understandable explanations for stance predictions using the datasets *Createdebate* and *Room For Debate* by building a stance reasoning tree using the rhetorical structure of the document and Dempster-Shafer theory (Dempster, 1967; Shafer, 1976).

Ding et al. (2024) used an LLM to generate a step-by-step explanation (a reasoning chain) for stance classification, which is then used as an additional training signal for a stance classifier. Moving forward towards these LLM capabilities, Weinzierl and Harabagiu (2024) introduced Tree-of-Counterfactual prompting to generate structured explanations for each possible stance label on the SemEval-2016 dataset. Taranukhin et al. (2024) employed *Chain-of-Thought* (*CoT*) explicit reasoning on the SemEval-2016 and P-Stance datasets, producing intermediate reasoning steps as explanations. Lan et al. (2024) applied a multi-agent debate framework (*COLA*) to the SemEval-2016 and VAST datasets, where different agents contribute partial explanations before a final prediction.

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)	(MAW) Mean Attention Weights	VAST	Token
Zhang et al. (2022)	ChatGPT	SemEval-2016 / P-Stance	Passage
Zhang et al. (2023)	Rule-entity binding	X Posts Dataset	Passage
Saha et al. (2024)	Argument-relevance weight	Createdebate / Room For Debate	Passage
Ding et al. (2024)	Chain-of-Thought (CoT)	SemEval-2016 / P-Stance / VAST	Passage
Weinzierl & Harabagiu (2024)	Tree-of-Counterfactual Prompting	SemEval-2016	Passage
Taranukhin et al. (2024)	CoT Explicit Reasoning	SemEval-2016 / P-Stance	Passage
Lan et al. (2024)	Multi-Agent Debate (COLA Framework)	SemEval-2016, VAST	Passage
Reveilhac & Schneider (2025)	Linguistic markers and rule-based model	SRQ (Stance in Replies and Quotes)	Passage
Muthusami et al. (2025)	Topic-guided transformers (BERTopic)	SemEval-2016	Passage
Nwaiwu et al. (2025)	LIME and Permutation Importance	FakeNewsChallenge	Token / Feature

Other contemporary methods incorporate linguistic and topical frameworks to achieve passage-level granularity. Reveilhac and Schneider (2025) used a rule-based model with linguistic markers on the *SRQ* dataset (Villa-Cox et al., 2020). Similarly, Muthusami et al. (2025) applied topic-guided transformers (*BERTopic*) to the SemEval-2016 dataset. Returning to this token and feature-level granularity in recent years, Nwaiwu et al. (2025) applied LIME (Ribeiro et al., 2016) and Permutation Importance to the FakeNewsChallenge dataset.

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 relate knowledge to linguistic theories related to pragmatics, discourse analysis, and sociolinguistics.

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 X posts
Trump vs. Hillary	Darwish et al. (2017)	Passage	3,450 X posts
Multi-target SD	Sobhani et al. (2017)	Passage	4,455 X posts
TW-BREXIT	Lai et al. (2020b)	Passage	∼5,400 X posts
P-Stance	Li et al. (2021)	Passage	21,574 X posts
ParlVote+	Abercrombie & Batista-Navarro (2022)	Passage	33,311 speeches
CoCoHD	Hiray et al. (2024)	Sentence	∼1,000 sentences
PolitiSky24	Rostami et al. (2025)	Passage	∼18M Bluesky posts
TruthStance	Ameen et al. (2026)	Passage	523,360 Truth Social posts

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 the English language (Table 3), we observed patterns in how stance has been studied:

- **Preference for social media posts:** A major preference for posts in social media platforms as the main material for stance datasets may reflect the easy availability and volume of data available through APIs or scraping of platforms like X, as well as the emerging Bluesky (Rostami et al., 2025), and Truth Social (Ameen et al., 2026), 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 rhetoric 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.
- **Preference for passage-level focus:** Excluding Hiray et al. (2024), datasets in English used for stance detection share the granularity of the passage level (multiple sentences).
- **Large sizes:** Again, excepting Hiray et al. (2024), most datasets use a large number of examples: from Trump vs. Hillary (3,450) to PolitiSky24 (∼18M), making their construction a formidable task. This extensive demand for examples is aligned with the need for *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.
- **Specificity:** In some cases, datasets for stance are built based on specific political events, eras, or issues, such as the 2016 U.S. Presidential Election in Trump vs. Hillary by Darwish et al. (2017), the Brexit referendum in TW-BREXIT by Lai, Patti, et al. (2020), and the fossil fuel consumption in *CoCoHD* (Hiray et al., 2024).
- **Preference for public opinion:** Most of the datasets for stance detection and classification, while comprehensive in their collection of public opinion in the social media *X*, *Bluesky*, or *Truth Social*, 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.

More recent work has begun to draw on parliamentary and congressional records for stance detection and classification (Abercrombie & Batista-Navarro, 2022; Hiray et al., 2024), but they either remain at the passage level (Abercrombie & Batista-Navarro, 2022) or focus on one specific political issue (Hiray et al., 2024).

Although single-sentence stance datasets and multi-sentence stance examples each offer distinct benefits for research and applications in NLP, the analysis of stance at the sentence level offers key benefits: (1) sentence-level stance usually consists 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 approach can help identify 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 LLMs (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) lack of research on stance classification using few-shot learning models – in general – and SetFit – in particular; (2) the vast majority of datasets for stance classification in English are entirely built from online debates on social media, and only two by professional politicians; (3) since most of the datasets for stance classification in English are built from social media posts or speeches, most of the studies of stance extend to the multi-sentence (passage) level; (4) lack of scholarly analysis on explainable stance classification in English at the sentence level using traditional linguistic features combined with neural models;

and (5) lack of investigation into the specific subjective stance expressions in English in a single sentence.

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 easier to interpret. For the sake of simplicity, we will refer mostly to subjective stance as "stance", using the complete name of *subjective 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 the 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. In this section, we describe the datasets and their preparation, including data collection, cleaning, and annotation (Section 3.1). We then explain how we set up StanceFeat and StanceFit models for the classification task (Section 3.2). Datasets, models, and code are available in our GitHub repository.

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 (Reyes, 2023) 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:

- The match of at least one political issue of 158 political issues (Appendix A) 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 367 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.").

- 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 was not less than seven tokens away. This automatic evaluation was made using spaCy's linguistic capabilities.
- 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), ensures the sentence has content with structured meaning.
- The sentence length with at least five tokens and at most 50 tokens, measured with spaCy's linguistic capabilities.
- The number of clauses is higher than zero but not more than three, measured with spaCy's linguistic capabilities.
- 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.
- Presence of singular or plural personal pronouns, "I", "me", "my", "mine", "myself", "we", "us", "our", "ours", "ourselves", automatically detected with spaCy's linguistic capabilities.

The filtering task resulted in a pool of 14,101 unlabeled sentences, which were analyzed through a bootstrapping approach of ten rounds of annotation and model inference, classifying the sentences into *support* or *oppose*. Initially (in iteration 1/10), we built our ground truth, a seed dataset (the seed of StanceSentences) with 180 sentences (90 supporting and 90 opposing), and a test dataset with 200 sentences (100 in each class). This test dataset remained strictly fixed and independent and never overlapped or interacted with any annotation or dataset expansion processes, serving consistently as the gold standard for evaluating each model iteration.

In each iteration, four annotators and the curator worked collectively, where the pool of sentences was submitted to StanceFit for inference, assigned each sentence a prediction confidence score that we used to create a ranking of sentences, from which only a batch of 100 sentences with perfect class balance with the highest score, representing the wide linguistic variety of subjective stance expressions, and a collective approval of the new candidates by human annotators in review meeting, were retained and added to StanceSentences dataset (as described in the Annotation Guidelines in Appendix B).

To ensure a methodological separation, at each iteration, the SetFit model was fine-tuned from scratch using only the incrementally expanded training set, validated against a validation split, and evaluated against the fixed test set. The validation split was used to monitor performance and guide decisions during each iteration, while the test set was used exclusively for final iteration reporting. The new sentences complied with a human review that followed these strict criteria:

1. Focus on sentences that convey the (individual or collective) speaker's (subjective) stance (like, "I believe that…", "We will foster…", "Our next step in the fight against…", "My administration has strengthened…"), not someone else's stance (like, "The president believes that…").
2. The stance target(s) had to be a clearly identified political issue (either present in the predefined list of political issues or not) and not personal pronouns or determiners ("this", "that", "those", "these", "he", "she", "his", "hers", "their", etc.). Each sentence in the dataset implicitly or explicitly references one or many issues, and stance classification decisions were always made in relation to these defined targets. Thus, our stance classification is inherently context-specific and target-dependent, avoiding the interpretation of stance as an isolated abstract notion.
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…". This preference aimed to allow easy finding of stance expressions, similar to those in Liu et al. (2023).
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 2). 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.

In every round, an anonymous review was conducted by two additional annotators, unaware of the initial classifications. The inter-annotator agreement (IAA) achieved a Fleiss's Kappa score of 0.709 on the stance classification facet, implemented using statsmodels (Seabold & Perktold, 2010), calculated among all annotators.

The annotators were students from a master's-level seminar who participated voluntarily without any financial compensation. They were thoroughly informed about the objectives of the study and the purpose of their annotations, and they consented to take part, driven by the desire to acquire practical experience and recognition in the dataset's publication. The names of the annotators are acknowledged in the publicly accessible dataset repository, ensuring that their contributions are properly recognized. Our project did not involve any sensitive data or vulnerable groups, which is why we did not pursue formal institutional approval.

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).

The sources we utilized fell into two categories: (1) publicly available websites that allowed for fair use and academic research, and (2) proprietary sites for which we obtained permission upon request. In both instances, we adhered to the conditions established by each source to guarantee the

Stance	Sentence
Support	And **I** firmly believe that excellence in education [TARGET] 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 [TARGET].
Support	**We** Republicans are agreed that full employment [TARGET] shall be a first objective of national policy.
Support	For almost 4 years, **my administration** has strengthened our military [TARGET].
Oppose	But **I** think that since there is no inflation in the economy, interest rates [TARGET] should not continue to go up.
Oppose	**Our** next step in the fight against crime [TARGET] is to take on gangs the way we once took on the mob.
Oppose	Recently **I** have said that NATO [TARGET] was obsolete, because it did not properly cover terror, and also, that many of the member countries were not paying their fair share.
Oppose	**We** have war against those who are using ideas in order to create extremism [TARGET] and threats towards the whole innocent people in the world.

Figure 2: Examples of Subjective Stance Sentences in the StanceSentences Dataset.

ethical and legal usage of their content, including proper citations and compliance with applicable legal and academic guidelines.

3.2. Experimental Set-Up

With the StanceSentences dataset built and verified, the following sections describe how we operationalized its content and configured the models used for stance classification. Our first step was the extraction of 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 Appendix C.) 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's performance and explainability. The construction of the LRBM was rooted on a linguistic foundation, following 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 the Bias Mitigation section.)

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, thesauruses, and querying ChatGPT 4 and 4o (OpenAI, 2024) when disambiguation was necessary.

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 • Oppose	• Support • Oppose
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" • "oppose"	• "support" • "oppose"
Splits/Instances	• Train: 972 (90% of 1080) • Validation: 108 (10% of 1080) • Test: 200	• Train: 1080 (84.38%) • Test: 200 (15.62%)
Stratification*	• Train: 486 s / 486 o • Validation: 54 s / 54 o • Test: 100 s / 100 o	• Train: 540 s / 540 o • Test: 100 s / 100 o
Metadata	• title (document) • source • semantic_frames	• title (document) • source • semantic_frames
Data Period	1939–2023	1939–2023

Note: (*) s = support, o = oppose.

We used the categories of stance features studied by Biber and Finegan (1989) (Table 1): *affect*, *evidentiality*, and *modality*. *Affect* is the range of expressed personal attitudes, which includes positive and negative *emotions*, *feelings*, *moods*, and general *mental states*. 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 *predictability* 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 features, 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 emphasize 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. Equally, 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 representativeness 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 lexicons and patterns of studied features can be found in the Appendix C.

This feature engineering modeled 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;
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:

Table 5: Linguistic Features Used to Build the StanceSentencesFeat Dataset.

Feature	POS	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
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

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.

We fit StanceFeat with StanceSentencesFeat using the training split (1080 sentence, or 84.38%) and the test split (200 sentences defined as "ground truth", or 15.62%), as shown in Table 4. The LR analysis of StanceFeat paved the way for a deeper investigation into the explainability of the neural model. Alongside this baseline, we configured a second, neural model – StanceFit – to enable direct comparison of classification behavior and feature importance across fundamentally different architectures.

We chose SetFit's special variation paraphrase-mpnet-base-v2 to create StanceFit since it aligns with the challenges of classifying political discourse texts, and it is especially adept at discerning the subtle differences in sentences that may convey similar messages with different wording. To fit StanceFit, we used StanceSentences, divided as follows: (1) our test split of 200 sentences, and (2) the 1080 sentences used to train StanceFeat, divided into 972 sentences (90%) as the train split, and 108 sentences (10%) as the validation split (Table 4).

To fine-tune StanceFit, we used 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).

3.3. Experiments

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:

$$\text{SHAP}_i = \sum_{j=1}^{n} \text{SHAP}_{ij} \tag{1}$$

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 1 illustrates the NLP explainability process using SHAP.

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.

3.4. Bias Mitigation

Bias mitigation efforts commenced during the feature selection process. We controlled thematic bias by focusing on specific POS (adjectives, adverbs, and verbs) since nouns – especially named entities – are frequently targets of stance. This control extended to the lexicon-building 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

This section presents the results of our analysis while discussing the feature engineering process and highlighting the key linguistic differences between stance expressions of *support* and *oppose* (Section 4.1). We then discuss the bootstrapping process to build StanceSentences (Section 4.2) and evaluate the performance of both models (Section 4.3). Lastly, we analyze StanceFit using SHAP through qualitative and quantitative methods (Section 4.4).

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 are 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.

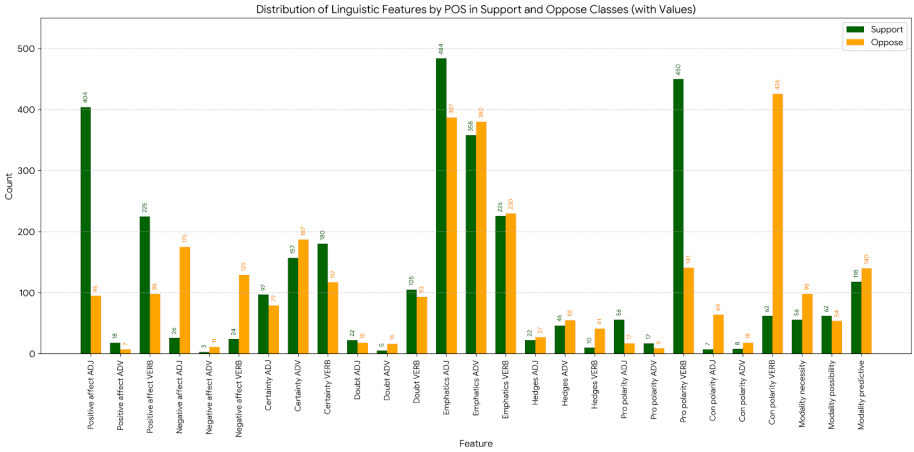

Figure 3: Distribution of Linguistic Features by POS in Support and Oppose Classes for Stance Detection in the StanceSentences dataset.

Figure 3 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

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

Table 7: LR Results for the Classification of Support and Oppose Stance Features.

Feature	β	$\mathbf{SE}_\beta$	z	p	$\mathbf{e}^\beta$
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
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

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.

in *oppose* stance. This finding underscores the alignment between affective language and stance, reflecting the emotional undertones embedded within pro and con arguments. The LR analysis (Table 7) provided insights into the importance of features, with pro polarity and positive affect as the strongest predictors for classifying sentences into the *support* class. Specifically, sentences containing pro-polarity expressions ($\beta = 1.867$, $e^{\beta} = 6.468$, $p < .001$) are approximately 6.5 times more likely to express a supportive stance compared to sentences without such features. Similarly, positive affect ($\beta = 1.777$, $e^{\beta} = 5.914$, $p < .001$) increases these odds nearly sixfold. Conversely, negative affect ($\beta = -1.980$, $e^{\beta} = 0.138$, $p < .001$) and con polarity ($\beta = -2.723$, $e^{\beta} = 0.065$, $p < .001$) significantly reduce the odds of sentences being classified as supportive, making them strong indicators of an opposing stance. Features such as certainty ($\beta = 0.138$, $p = .435$), emphatics ($\beta = 0.033$, $p = .889$), and doubt ($\beta = 0.309$, $p = .164$) were statistically insignificant, while hedges ($\beta = -0.321$, $p = .097$) had only a weak, non-significant tendency towards opposing stance.

4.2. Bootstrapping Process

Table 8 shows the metrics of StanceFit along the bootstrapping process to build the StanceSentences dataset. 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 were added, the rate of improvement in model performance tended to plateau, something clearly visible in the latter iterations, suggesting that the model extracted as much generalizable knowledge as it could from the data provided.

Across the iterations, approximately one in three sentences proposed by StanceFit was accepted by annotators during the review meetings, which helped to correct and diversify the dataset. In general, the data analysis suggests that StanceFit adapted well to incremental data in a bootstrap fine-tuning framework. To visually evaluate the adaptation of the model to the classification task, Figure 4 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 fine-tuning 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, a 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. The fine-tuning process in iteration 10 took 6 hours and 53 minutes.

The evaluation embeddings validate how well the model fits 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 fine-tuning 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. Datasets and models are publicly available (Reyes, 2026).

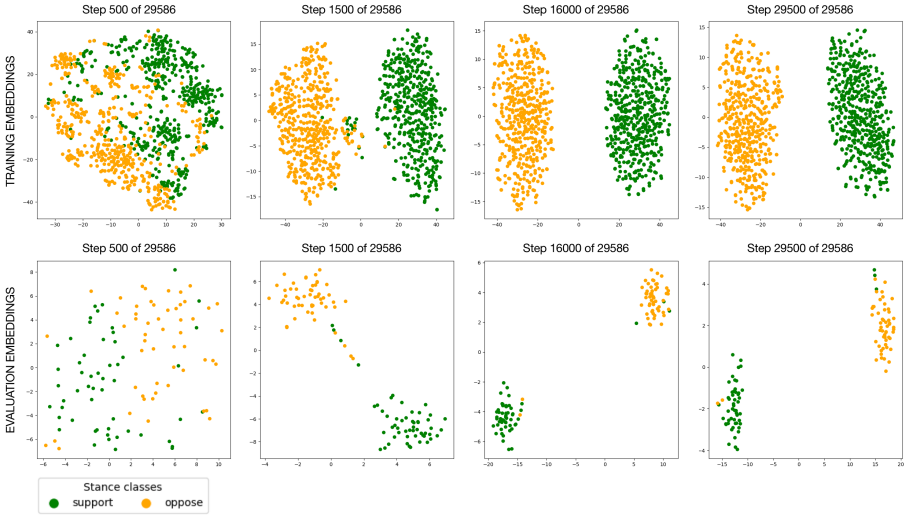

Figure 4: t-SNE plots of Iteration 10 of the Bootstrapping Fine-tuning for Stance Classification.

4.3. Models Performance

As seen in Table 9, StanceFit achieves near-perfect scores (0.995) 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 (*)	$\begin{pmatrix} 84 & 17 \\ 21 & 79 \end{pmatrix}$	$\begin{pmatrix} 100 & 0 \\ 1 & 99 \end{pmatrix}$
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. Rows: predicted, Columns: actual.
(1) Across-class metrics are macro, and class-wise metrics are not averaged.

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 not have been explicitly clear but are implied through paraphrasing or similar expressions.

4.4. 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.

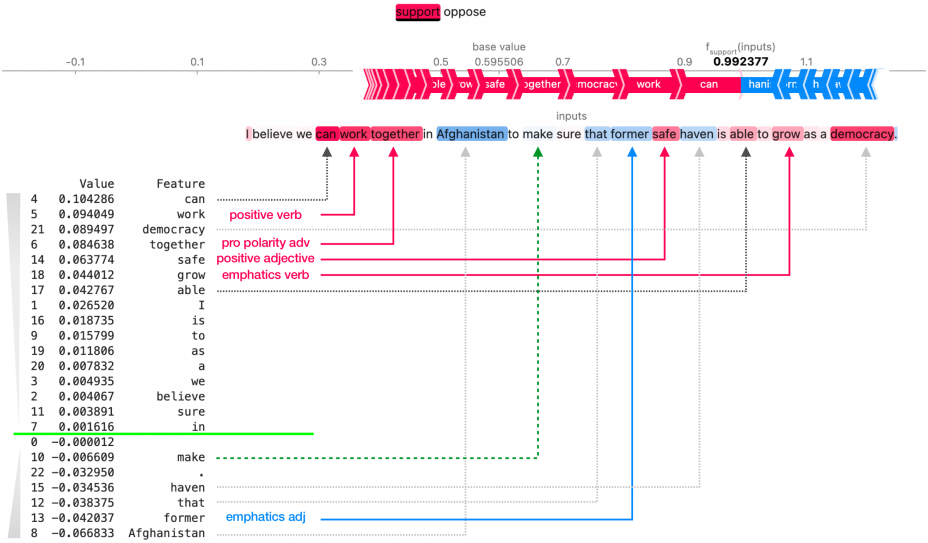

Figure 5: Example 1, SHAP explanation on an example of the Support Class and Alignment to Lexicons in the LRBM.

In Example 1 (Figure 5), 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.

This analysis follows the colors of the arrow lines in Figure 5:

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 C, 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 C, 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 6), 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). Through this example, we can observe other complementary behaviors in our analysis of StanceFit:

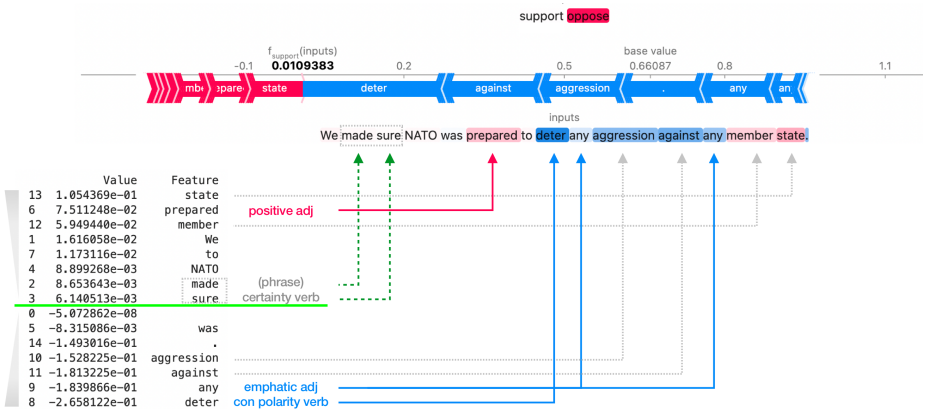

Figure 6: Example 2, SHAP explanation on an example of the Oppose Class and Alignment to Lexicons in the LRBM.

1. Support term in lexicons (red): The adjective "prepared" has the highest 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 highest 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 C, 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 makes StanceFit consider it irrelevant in the *oppose* class.

The comparative analysis between StanceFeat's LR coefficients (β) 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 β 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 β and 0.930 as SHAP value. Negative affect and con polarity show significant negative contributions, with negative affect having normalized values of -1.235 β and -0.827 as SHAP value, and con polarity exhibiting the strongest negative influence with values of -1.730 β and -2.011 as SHAP value. These negative contributions are crucial as they influence the *oppose* class predictions. Certainty exhibits a modest positive effect, β 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, β at 0.210 and SHAP values at -0.172, highlighting its role in decreasing the likelihood of *support*. Figure 7 shows how both models rely

Table 10: Comparison of Feature Importance in **StanceFeat**'s LR Coefficients and **StanceFit**'s SHAP values.

Feature	**Raw**		**Normalized**	
	LR β	SHAP Values	LR β	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

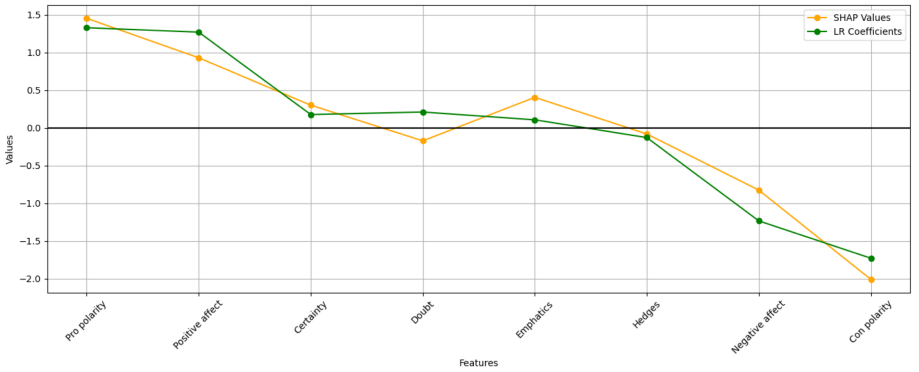

Figure 7: *Comparison of normalized SHAP values and Logistic Regression Coefficients.*

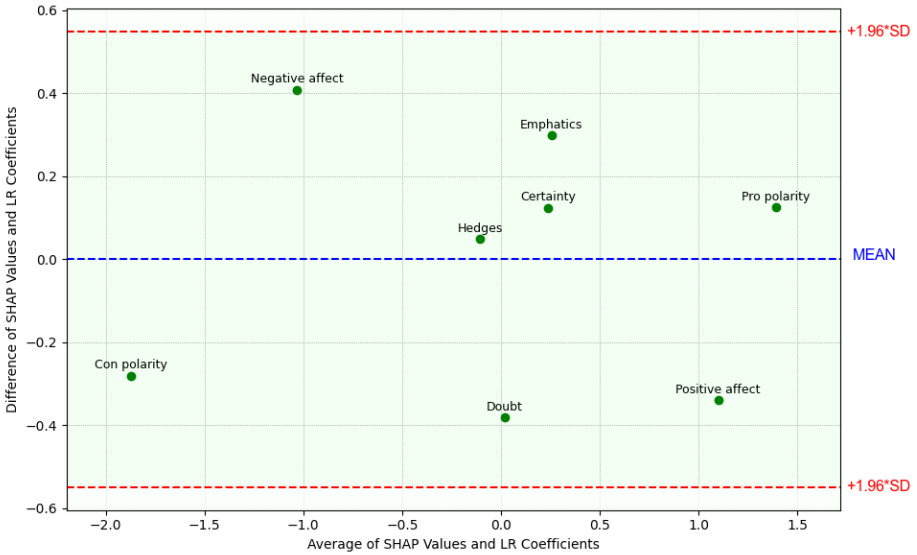

Figure 8: Bland-Altman Plot of Comparison of SHAP values and Logistic Regression Coefficients.

on the studied linguistic features similarly, allowing us to validate through this dual-analysis how linguistic features impact 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.

A two-sample Welch's t-test was conducted to compare the means of LR β 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 8 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. Oppositely, 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. Conclusion

The StanceSentences dataset may be a foundation 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 studies. 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 pertinence 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 safely. This claim is observable in Figure 3, in which certainty and emphatics have 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 studies *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.

References

Abercrombie, G., & Batista-Navarro, R. (2022). Policy-focused stance detection in parliamentary debate speeches. *North European Journal of Language Technology (NEJLT)*, *8*(1). https://doi.org/10.3384/nejlt.2000-1533.2022.3385

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, 2623–2631. https://doi.org/10.1145/3292500.3330701

AlDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. https://doi.org/10.48550/arXiv.2006.03644

Allaway, E., & McKeown, K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8913–8931. https://doi.org/10.18653/v1/2020.emnlp-main.717

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

Ameen et al. (2026). Truthstance: An annotated dataset of conversations on truth social [arXiv preprint]. https://doi.org/10.48550/arXiv.2602.14406

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. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, 1–9. https://aclanthology.org/W11-1701

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The longman grammar of spoken and written english*. Longman.

Chauhan, D., Kumar, R., & Ekbal, A. (2019). Attention based shared representation for multi-task stance detection and sentiment analysis. *International Conference on Neural Information Processing (ICONIP)*, 661–669. https://doi.org/10.1007/978-3-030-36802-9_70

Chen, P., Ye, K., & Cui, X. (2021). Integrating n-gram features into pre-trained model: A novel ensemble model for multi-target stance detection. *International conference on artificial neural networks*, *12893*, 269–279. https://doi.org/10.1007/978-3-030-86365-4_22

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., & Roth, D. (2019). Seeing things from a different angle: Discovering diverse perspectives about claims. *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

Chilton, P. (2004). *Analyzing political discourse: Theory and practice*. Routledge.

Darwish, K., Magdy, W., & Zanouda, T. (2017). Trump vs. Hillary: What went viral during the 2016 US presidential election. https://doi.org/10.48550/arXiv.1707.03375

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, *38*(2), 325–339. https://doi.org/10.1214/aoms/1177698950

Ding, D., Dai, G., Peng, C., Peng, X., Zhang, B., & Huang, H. (2024). Distantly supervised explainable stance detection via chain-of-thought supervision. *Mathematics*, *12*(7), 1119. https://doi.org/10.3390/math12071119

Du, J., Xu, R., He, Y., & Gui, L. (2017). Stance classification with target-specific neural attention networks. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3988–3994. https://doi.org/10.24963/ijcai.2017/557

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.

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. *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

Gómez-Suta, M., Echeverry-Correa, J., & Soto-Mejía, J. A. (2023). Stance detection in tweets: A topic modeling approach supporting explainability. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2022.119046

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. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 1859–1874.

Hasan, K. S., & Ng, V. (2013). Frame semantics for stance classification. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 124–132. https://aclanthology.org/W13-3514

Hiray, S., et al. (2024). Cocohd: Congress committee hearing dataset. *Findings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15529–15542. https://doi.org/10.18653/v1/2024.findings-emnlp.911

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). Spacy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jayaram, V., & Allaway, E. (2021). Human rationales as attribution priors for explainable stance detection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5540–5554. https://doi.org/10.18653/v1/2021.emnlp-main.450

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

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

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

Lai, M., Patti, V., Ruffo, G., & Rosso, P. (2020). 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

Lan, X., Gao, C., Jin, D., & Li, Y. (2024). Stance detection with collaborative role-infused llm-based agents. *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media*, *18*(1), 891–903. https://doi.org/10.1609/icwsm.v18i1.31360

Li, Y., & Caragea, C. (2019). Multi-task stance detection with sentiment and stance lexicons. *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. https://doi.org/10.18653/v1/D19-1657

Li, Y., Wang, S., Lin, C., Guerin, F., & Barrault, L. (2023). Framebert: Conceptual metaphor detection with frame embedding learning. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1558–1563. https://doi.org/10.18653/v1/2023.eacl-main.114

Liu, Z., Yap, Y. K., Chieu, H. L., & Chen, N. (2023). Guiding computational stance detection with expanded stance triangle framework. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3987–4001). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.220

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in english*. Palgrave Macmillan.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). A dataset for detecting stance in tweets. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3945–3952.

Mohtarami, M., Baly, R., Glass, J., Nakov, P., Marquez, L., & Moschitti, A. (2018). Automatic stance detection using end-to-end memory networks. *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. https://doi.org/10.18653/v1/N18-1070

Muthusami, R., Saritha, K., Rao, K. S., Sugapriya, P., & Saveetha, G. (2025). Interpretable stance detection in social media via topic-guided transformers. *Discover Artificial Intelligence*, *5*(1), 355.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019). Explanation of machine learning models using improved shapley additive explanation. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 546. https://doi.org/10.1145/3307339.3343255

Nwaiwu, S., Jongsawat, N., & Tungkasthan, A. (2025). Decoding disinformation: A feature-driven explainable ai approach to multi-domain fake news detection. *Applied Sciences*, *15*(17), 9498. https://doi.org/10.3390/app15179498

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

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., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*. https://doi.org/10.48550/arXiv.1201.0490

Peters, G., & Woolley, J. T. (n.d.). The american presidency project. https://www.presidency.ucsb.edu/

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2019). Stancy: Stance classification based on consistency cues. *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. https://doi.org/10.18653/v1/D19-1675

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. https://arxiv.org/abs/1908.10084

Reveilhac, M., & Schneider, G. (2025). Evaluating a transparent and interpretable approach to stance detection using linguistic markers in social media data. *International Journal of Corpus Linguistics*, *30*(2), 195–233. https://doi.org/10.1075/ijcl.24132.rev

Reyes, J. F. (2023). Webcrawler, a web crawler for political discourse texts [GitHub repository. Source code]. https://github.com/pacoreyes/webCrawler

Reyes, J. (2026). Hugging face models and datasets. https://huggingface.co/pacoreyes

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. https://doi.org/10.48550/arxiv.1602.04938

Rostami, P., Rahimzadeh, V., Adibi, A., & Shakery, A. (2025). Politisky24: U.s. political bluesky dataset with user stance labels. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 21976–21993. https://doi.org/10.18653/v1/2025.findings-emnlp.1198

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.

Saha, R. R., Lakshmanan, L. V. S., & Ng, R. (2024). Stance detection with explanations. *Computational Linguistics*, *50*(1), 193–235. https://doi.org/10.1162/coli_a_00501

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 57–61. https://doi.org/10.25080/Majora-92bf1922-011

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press. https://doi.org/10.1515/9780691214696

Sobhani, P., Inkpen, D., & Zhu, X. (2017). A dataset for multi-target stance detection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 551–557.

Sobhani, P., Mohammad, S., & Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 159–169. https://doi.org/10.18653/v1/S16-2021

Somasundaran, S., & Wiebe, J. (2010). Recognizing stances in ideological on-line debates. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124.

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

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. https://doi.org/10.1007/978-3-319-50496-4_76

Sun, Q., Wang, Z., Zhu, Q., & Zhou, G. (2018). Stance detection with hierarchical attention network. *Proceedings of the 27th International Conference on Computational Linguistics*, 2399–2409.

Taranukhin, M., Shwartz, V., & Milios, E. (2024). Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1326–1337. https://aclanthology.org/2024.lrec-main.1326/

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

Tunstall, L., Reimers, N., Seo Jo, U. E., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. https://arxiv.org/abs/2209.11055

Villa-Cox, R., Kumar, S., Babcock, M., & Carley, K. M. (2020). Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. https://doi.org/10.48550/arXiv.2006.00691

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

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.

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

Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wasserblat, M. (2021, December). Sentence transformer fine-tuning (setfit). https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Sentence-Transformer-Fine-Tuning-SetFit/post/1407712

Weinzierl, M. A., & Harabagiu, S. M. (2024). Tree-of-counterfactual prompting for zero-shot stance detection. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–880. https://doi.org/10.18653/v1/2024.acl-long.49

Wilson, J. (1990). *Politically speaking: The pragmatic analysis of political language*. Basil Blackwell.

Zhang, B., Ding, D., Xu, G., Guo, J., Huang, Z., & Huang, X. (2023). Twitter stance detection via neural production systems. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094597

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

A. List of Political Issues

1. Exact match

2019-nCoV, *COVID*, *COVID-19*, *EPW*, *GDP*, *GLBT*, *I4.0*, *LGB*, *LGBT*, *N.A.T.O.*, *NATO*, *POWR*, *PAS*, *SARS-CoV-2*, *U.N.*, *UAS*, *UAV*, *UBI*, *UN*, *UNO*

2. Hyphenated terms

computer-orient crime, *health-care*, *house-building*, *ill-house people*, *ill-housed person*, *non-heterosexual*, *non-neoplastic*, *same-sex marriage*

3. Lemmatized terms

4th Industrial Revolution, *aboriginal people*, *aborigines*, *abortion*, *abuse of substance*, *Afghanistan*, *agricultural holding*, *agriculture*, *ailment*, *alcohol use disorder*, *alcoholism*, *American Indians*, *Amerindians*, *anticonception*, *arm control*, *armed force*, *army*, *atomic energy*, *automation*, *bank industry*, *banking*, *banking industry*, *basic income*, *bi guarantee*, *big pharma*, *birth control*, *birth rate*, *birthrate*, *border*, *cannabis*, *China*, *citizen income*, *civil right*, *climate change*, *climatic variation*, *compensation*, *computer crime*, *constitution*, *contraception*, *Corea*, *coronavirus*, *correctional facility*, *correctional institution*, *corruption*, *coup*, *coup d'état*, *crime*, *criminality*, *custom duty*, *cyber warfare*, *cybercrime*, *democracy*, *deportation*, *depression*, *dictatorship*, *dictature*, *disability*, *discrimination*, *disease*, *drone*, *drug*, *drug abuse*, *ecology*, *edible*, *educate*, *education*, *embargo*, *employment*, *energy*, *epidemic*, *ethnic conflict*, *ethnic extremism*, *ethnic hatred*, *ethnic prejudice*, *ethnic struggle*, *ethnic tension*, *ethnic war*, *euthanasia*, *family planning*, *farm*, *firearm control*, *food*, *foreign policy*, *foreign relation policy*, *Fourth Industrial Revolution*, *frontier*, *gay marriage*, *gender*, *gerrymander*, *gerrymandering*, *glbt people*, *glbt person*, *global development*, *globalisation*, *globalism*, *globalization*, *gradual disarmament*, *gross domestic product*, *gross sexual imposition*, *gun ban*, *gun control*, *gun law*, *gun prohibition*, *gun regulation*, *handicap*, *harmful use*, *hate speech*, *health care*, *health coverage*, *health insurance*, *health profession*, *health service*, *healthcare*, *high price*, *home building*, *home construction*, *homeless*, *homeless person*, *homosexual marriage*, *hostility*, *house building*, *house construction*, *house production*, *housing*, *human migration*, *human race*, *human right*, *human smuggling*, *human trafficking*, *humanitarian aid*, *humanitarian assistance*, *humanitarian relief*, *illegal drug*, *illegal immigration*, *illegal migration*, *illegal substance*, *illicit substance*, *illness*, *indigenous Americans*, *indigenous culture*, *indigenous people*, *indigenous people of the Americas*, *industry 4*, *industry 4.0*, *infectious agent*, *inflation*, *infrastructure*, *international affair*, *international aid*, *international development*, *international relation*, *investment*, *Iran*, *jail*, *joblessness*, *Korea*, *law project*,

lesbian marriage, *liberalism*, *lobby*, *malpolitics*, *marihuana*, *marijuana*, *marriage equality*, *mass medium*, *mass public shooting*, *mass shooting*, *mass surveillance*, *media*, *medical care*, *medical insurance*, *merger and acquisition*, *MeToo*, *middle class*, *migration*, *migratory movement*, *military*, *military affair*, *military alliance*, *military budget*, *military expenditure*, *military housing*, *military spending*, *military technology*, *minimum pay*, *minimum wage*, *monetary erosion*, *mortgage*, *mortgage loan*, *narcotic*, *national security*, *nationalism*, *nationalist ideology*, *Native Americans*, *native people*, *natural calamity*, *natural disaster*, *natural hazard*, *net neutrality*, *network neutrality*, *North Atlantic Alliance*, *North Atlantic Treaty Organisation*, *North Atlantic Treaty Organization*, *nuclear energy*, *nuclear power*, *outbreak*, *pandemic*, *peculate*, *peculation*, *penitentiary*, *pension*, *people flow*, *People Republic of China*, *pharma*, *pharmaceutical industry*, *pilotless aircraft*, *plan parenthood*, *planning of family*, *plastic*, *polarisation*, *polarization*, *political corruption*, *political freedom*, *political polarization*, *political turmoil*, *political violence*, *pollution*, *populism*, *populist*, *poverty*, *pregnancy prevention*, *pressure group*, *price*, *price rise*, *prison*, *prisoner of war*, *PsW*, *public health*, *public hygiene*, *race conflict*, *race struggle*, *race war*, *racial conflict*, *racial hatred*, *racial intolerance*, *racial prejudice*, *racial segregation*, *racial tension*, *racism*, *rape*, *real estate*, *real property*, *recession*, *refugee*, *relief work*, *religion*, *religious tradition*, *remotely pilot aircraft system*, *remuneration*, *residential construction*, *resignation*, *retirement plan*, *retribution*, *richness*, *Russia*, *salary*, *same sex marriage*, *scarcity*, *school violence*, *science*, *segregation*, *sex education*, *sexual and gender minority*, *sexual assault*, *shooting spree*, *shortage*, *sickness*, *social conflict*, *social hygiene*, *social inequality*, *social insurance*, *social medium*, *social program*, *social protection*, *social security*, *social security contribution*, *socioeconomic conflict*, *sociopolitical conflict*, *spree shooting*, *stimulus*, *street people*, *substance abuse*, *supremacism*, *surveillance*, *sustenance*, *Syria*, *takeover*, *tariff*, *tax*, *tech*, *technology*, *termination of pregnancy*, *terrorism*, *the press*, *tolerance*, *toleration*, *total reward*, *trade*, *trading*, *traditional people*, *transaction*, *Ukraine*, *uncrewed aerial vehicle*, *unemployment*, *unhouse people*, *unhouse person*, *United Nations*, *universal basic income*, *universal demogrant*, *unmanned aerial device*, *unmanned aerial system*, *unmanned aerial vehicle*, *unmanned aircraft*, *vaccination*, *vaccine*, *vet*, *veteran*, *vigilance*, *violence in politic*, *virus*, *wage*, *war*, *wealth*, *weapon control*, *web neutrality*, *welfare*, *welfare payment*, *woman*

B. Annotation Guidelines of Dataset 1: Support and Oppose Stance

Author: [Juan-Francisco Reyes]

This is an adapted version of the original annotation procedure, retaining content relevant to this paper

This document delineates the process of annotating a dataset to train a stance classification model using SetFit. The objective is to build the dataset using the bootstrapping approach, a semi-supervised learning approach where a small set of labeled data (seed dataset) is used to train a preliminary model, which then annotates more data. After validation, this newly annotated data is added to the training set, and the process iterates. The primary goal of bootstrapping in this project is to expand a dataset efficiently, starting from a small, manually labeled dataset.

Task Using the *seed* dataset (provided by the lecturer), increase it iteratively with representative new sentences, with partial batches of 100 sentences each per round.

Stance A political stance, also known as a political position or viewpoint, is an individual's or group's set of beliefs, opinions, and values regarding politics and governance. It reflects how one thinks about various political issues, policies, and ideologies.

Stance classification typically operates by categorizing the stance as either in support (or favor) or oppose (or against), regarding the target.

1. **Support stance:** The speaker shows agreement, approval, or positive feelings towards the target. For example, "I fully support renewable energy initiatives". Here, the *support* stance is towards renewable energy initiatives. Words often used: "agree", "love", "support", "advocate", "praise".
2. **Oppose stance:** The speaker expresses disagreement, disapproval, or negative feelings. Example: "I am against deforestation". The *oppose* stance is towards deforestation. Common words: "disagree", "hate", "oppose", "criticize", "condemn".

Target The (stance) target refers to the specific subject, entity, or idea that a speaker or writer is expressing their attitude, feelings, or viewpoint towards. To identify the target in a piece of communication, look for what or whom the opinions, feelings, or attitudes are being directed at. It is usually easily found as the sentence's subject. In our project, the target must be a political issue (see list of political issues).

Selection Criteria In this project, we will focus on sentences with explicit stance and sometimes of moderate complexity of stance, excluding highly implicit expressions charged with euphemisms, metaphors, and other figurative or rhetorical devices to make the statement less obvious.

Opinions are usually a clear expression of stance. The identification of the opinion may rely on positive or negative adjectives or verbs that leave no doubt about the stance.

- Example of *oppose* stance: "And I think the threats of terrorism and the hatred that presently exists, the threat of war, the threat of economic boycotts and punishment against Egypt, are certainly not conducive to realizing the hopes of the Palestinian people". Target: "terrorism".
- Example of *support* stance: "We've made real, continuing investments in science and technology, which I think are pivotal to the long-term health of the economy and the continuation of this productivity increase". Target: "economy".

Acceptable stance expressions should comply with the following four criteria:

- **Subjective stance:** individual or collective, like in, "I believe that…" or "we reject…", not someone else's stance, like in, "The president believes that…"
- **Identifiable target:** must be clearly identified political issues (one or many) and not personal pronouns or determiners ("this", "that", "those", "these", "he", "she", "his", "hers", "their", etc.), like in "This is the most promising…"
- **Explicit stance:** like in "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 for…" or "we ought to be a little tougher on". You should avoid 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."

Step 1: Access the Annotation Tool Use the hyperparameter optimization script (provided by the lecturer in a Google Colab document) using Optuna and the dataset split (into three JSONL files: train and test), also provided by the lecturer. Then, fine-tune the seed model. You are supposed to get an accuracy score near 0.960 in the first round and avoid declining that metric. Save the model according to SetFit's official documentation.

Step 2: Apply the Seed Model to Unlabeled Data The fine-tuned model is then applied to a larger set of unlabeled data. The goal here is to make predictions or annotations on this data, effectively generating new, albeit potentially noisy, labeled instances. Download segments of examples available when necessary.

Step 3: Select High-Confidence Predictions From these predictions, a subset is chosen based on confidence criteria from the model predictions with high probability.

Step 4: Attend the Review Meetings In the review meetings, we collectively review new candidate sentences, rejecting or adding new candidates if necessary. Collect a balanced representation of both classes (50 sentences per class).

Step 5: Aggregate New Examples to the Dataset Merge new examples with the train dataset.

Step 6: Fine-tune the Model with the Increased Dataset Fine-tune the model and test it using the test dataset. If the performance of the model decreases, review the aggregated subset, organizing an anonymous review among annotators. When the model performance increases, repeat Step 2.

Note on Bootstrapping The main challenge to consider with bootstrapping is that the model might reinforce its own errors (confirmation bias), especially if the high-confidence predictions are not accurate. Also, the diversity of the dataset might not increase significantly, as the model tends to predict labels similar to those in the initial training set.

Inter-Annotator Agreement (IAA) (The IAA was followed by different annotators who did not participate in the bootstrapping process.)

The objective is to ensure a high degree of agreement in classifying (into *support* and *oppose*). This IAA assessment will aim to verify the consistency and reliability of the annotations, which is crucial for the subsequent analysis involving the model training. You are supposed to annotate the agreement assessment based on the following four facets:

1. Correct class: The classification is correct.
2. Clear target: The stance target is clear and recognizable (see list of political issues).
3. Subjective stance expression: The stance belongs to the speaker, not anyone else's.
4. Explicit stance expression: No ambiguities detected.

In the shared Google Spreadsheet, add a cross, "X", whenever a discrepancy is found. Leave the cell blank if no discrepancy was found. If a discrepancy is found, it should be discussed in the weekly sessions. Attend the review meetings in the weekly sessions to discuss disagreements and clarify any ambiguities.

C. Lexicons and Linguistic Features

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 adjective]

11. Doubt adverbs

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

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, threat*]

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, threat*]

[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

Correspondence

Juan-Francisco Reyes

Brandenburgische Technische Universität Cottbus-Senftenberg
Fakultät 1 – Institut für Informatik
Cottbus, Germany
pacoreyesp@gmail.com