# Stance Detection with Explanations

Rudra Ranajee Saha
University of British Columbia
Department of Computer Science
rrs99@cs.ubc.ca

Laks V. S. Lakshmanan
University of British Columbia
Department of Computer Science
laks@cs.ubc.ca

Raymond T. Ng
University of British Columbia
Department of Computer Science
rng@cs.ubc.ca



*Identification of stance has recently gained a lot of attention with the extreme growth of fake news and filter bubbles. Over the last decade, many feature-based and deep-learning approaches have been proposed to solve stance detection. However, almost none of the existing works focus on providing a meaningful explanation for their prediction. In this work, we study stance detection with an emphasis on generating explanations for the predicted stance by capturing the pivotal argumentative structure embedded in a document. We propose to build a stance tree that utilizes rhetorical parsing to construct an evidence tree and to use Dempster Shafer Theory to aggregate the evidence. Human studies show that our unsupervised technique of generating stance explanations outperforms the SOTA extractive summarization method in terms of informativeness, non-redundancy, coverage, and overall quality. Furthermore, experiments show that our explanation-based stance prediction excels or matches the performance of the SOTA model on various benchmark datasets.*

## 1. Introduction

Schiller, Daxenberger, and Gurevych (2021) defined the task of stance detection (SD) with two inputs—a topic of discussion and an author's comment towards that. In the existing literature, the topic can vary from a single phrase (also known as *target*) like "climate change" to a well-formed sentence like "Climate change is not real." Similarly, the author's comment can vary from a short tweet to a long essay (Faulkner 2014). Stance labels can be one of $\{pro, con, neutral, unrelated\}$. While there have been a plethora of works on target-based SD (Siddiqua, Chy, and Aono 2019; Du et al. 2017; Zhou,

Action Editor: Saif M. Mohammad. Submission received: 18 April 2023; revised version received: 29 September 2023; accepted for publication: 25 October 2023.

https://doi.org/10.1162/coli_a_00501

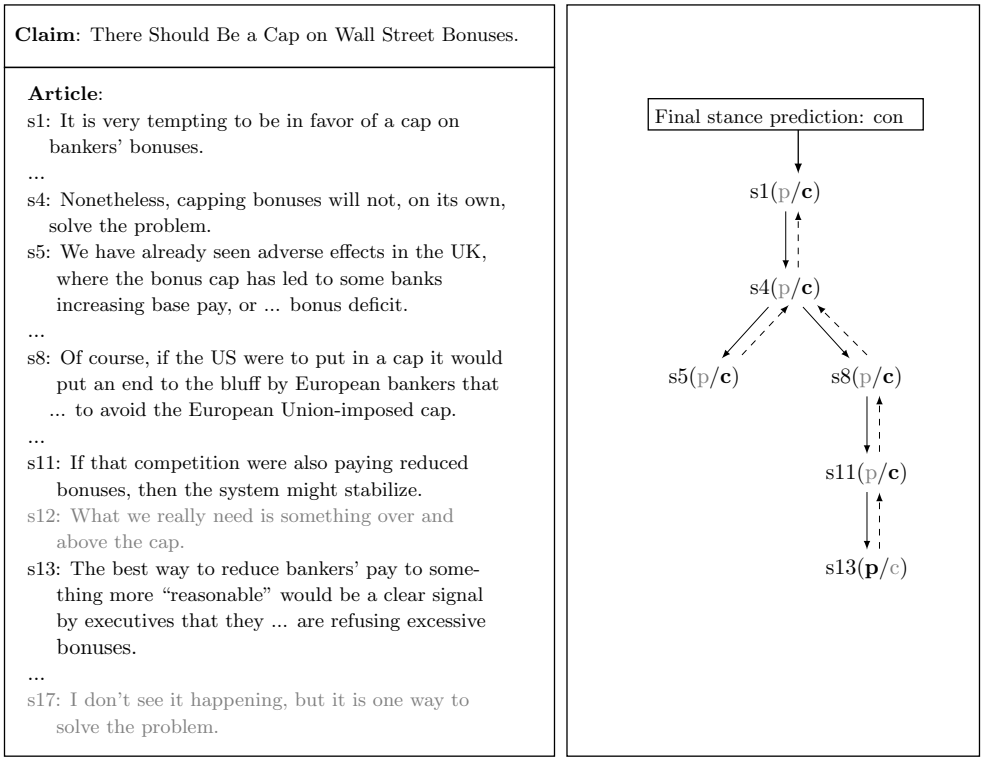



**Figure 1**
*Left*: A sample claim at the top followed by an article, decomposed into sentences (s1-s17) with index number; we show a subset of sentences due to the lengthiness of the article; sentences that are not part of the stance tree, i.e., pruned, are colored in gray; *Right*: Stance tree for the claim and article in the left; each node refers to one of the sentence indices in the article on the left; possible stance predictions for each sentence are shown inside the parentheses (*p* for pro, *c* for con); The prediction made by the model is shown in bold; the final verdict is shown at the top of the tree. The dashed arrows refer to the prediction propagation from bottom to top of the stance tree, while the solid arrows represent the explanation flow, extracted by our approach.

Cristea, and Shi 2017; Wei, Mao, and Zeng 2018), the claim-article setting of SD has received less attention. Figure 1 (left), gives an example of a claim "There should be a cap on Wall Street bonuses.", and the task is to determine the stance of the article (i.e., s1–s17) toward the claim and extract the most crucial arguments from the article that justifies the stance. The claim-article setting poses more challenges than the conventional, more researched, topic-tweet and topic-sentence settings. First, a claim, unlike a topic, conveys a stance of its own that needs to be analyzed. Second, articles tend to be dense on opinions and/or facts, often combining them with justifications.

The majority of the recent approaches for SD make use of large language models, trained using deep neural nets (Fang et al. 2019; Schiller, Daxenberger, and Gurevych 2021). A crucial aspect of SD that has been under-studied is the *explainability* of the predicted stance. A popular strategy to expose the "knowledge" of the large models is known as *probing*. However, instead of generating task-specific explanations, probing only provides abstract information, for example, which layers of a large model learn

syntax or semantic representation, which is not insightful as an explanation for SD. Another possible strategy is to extract the most important phrases from a document using a trained model via attention. An attention-based explanation is often non-contextualized and semantically incomplete and is hardly sufficient when we focus on articles or essays. Only by recognizing the argumentative structure can we hope to fully realize the article author's reasoning for taking a stance.

In this article, we propose a structure called the **stance tree**, which captures the argumentative structure hidden in an article in the form of a discourse tree. Rhetorical parsing (RST) (Mann and Thompson 1988) is well studied and is used for various downstream tasks like argument extraction (Stab and Gurevych 2017; Hewett et al. 2019), summarization (Pu, Wang, and Demberg 2023; Hou and Lu 2020), sentiment analysis (Qian et al. 2022; Kraus and Feuerriegel 2019), and machine translation (Liu, Shi, and Chen 2020; Wu et al.). Based on document segmentation, RST theory proposes a hierarchical discourse structure, called **discourse parse tree** (DPT), where the leaves are clause-like units of the document, known as **elementary discourse units** (EDUs). Adjacent EDUs are connected via internal nodes to form the hierarchy. To create a more compact representation of arguments, we formulate a **discourse dependency tree** (DDT), derived from the DPT, to better map the dependencies between EDUs (Li et al. 2014; Hirao et al. 2013). We propose a framework (see subsection 4.1) which allows the conversion from an EDU-based DPT (resp., DDT) to a sentence-based DPT (resp., DDT). We use a stance predictor model to make predictions for each EDU (resp., sentence) in the stance tree and combine the predictions using Dempster Shafer Theory (DST) (Shafer 1976), which has been widely applied for decision-making, to aggregate evidence to reach a final conclusion. We also build a pruning module to discard irrelevant arguments as well as a (dis)agreement detection module for a better understanding of (dis)agreement in SD. As an illustration of these ideas, Figure 1 (Left) shows a claim and an article and Figure 1 (Right) shows the corresponding stance tree. The final stance label, appearing at the root node is *con*. Sentences *s12* and *s17* are not part of the stance tree as explained in subsection 4.4 and subsection 4.5. The predicted stance for each sentence (subsection 4.3) is shown in bold and the propagation of predictions is shown in dotted arrows in Figure 1 (Right). We explain the application of DST in our proposed stance tree in detail in Figure 2.

One of the main contributions of this article is the identification of stance explanations in the form of a structured collection of arguments from the article. Based on RST, stance tree organizes the arguments in an article in a hierarchical fashion, where more relevant arguments occupy the top levels in the tree. This enables us to prepare the explanation in a completely unsupervised fashion under a user-specified budget if any. For example, we can extract the explanations from Figure 1 (Right) by looking at the sentences appearing closer to the root. For example, for a two-line explanation, sentences *s1* and *s4* are the best candidates. An additional noteworthy feature of our proposed architecture is the incorporation of rhetorical parsing and the unsupervised nature of explanation generation. This unique combination allows seamless integration with any stance detection model, setting our work apart from existing approaches (Du et al. 2017; Li and Caragea 2019; Popat et al. 2019) where explanation modules are often architecture-specific. This adaptability and versatility further enhance the practicality and applicability of our stance explanation approach, paving the way for more informed and comprehensive analysis in stance detection tasks across various datasets and applications.

By revealing the pivotal arguments from an article, this work aims to benefit various individuals and groups, such as researchers and academics, journalists and

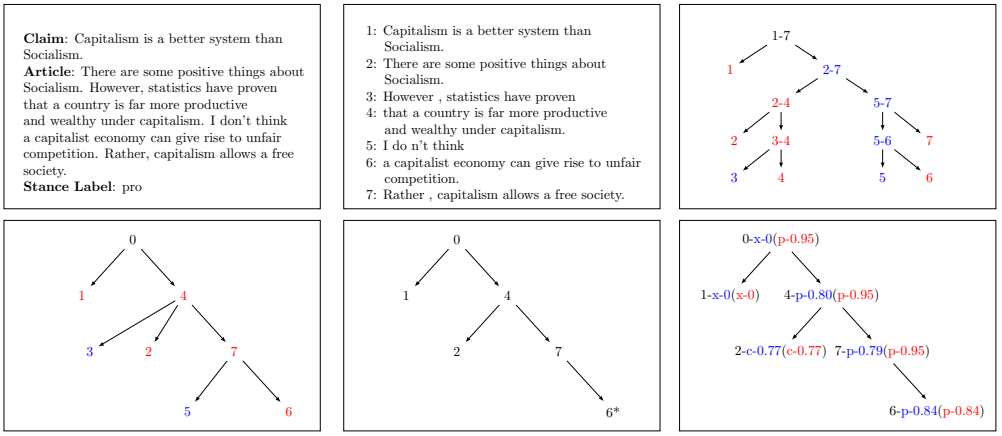



**Figure 2**
Various stages of our pipeline: **top row –** *Left*: A claim, article, and the stance label; *Middle*: Discourse segmented (EDU level); *Right*: Discourse Parse Tree, red nodes are the nucleus, blue nodes are satellite; **bottom row –** *Left*: Discourse Dependency Tree; *Middle*: Pruned Discourse Dependency Tree; *Right*: Evidence Aggregation, the blue part is prediction and confidence from stance predictor, the red part is combined prediction and confidence using DST theory.

fact-checkers, students and educators, debaters and public speakers, and content creators. Some common purposes served by such explanations are the following: (i) **Interpretability and Transparency**: The explanation helps users understand how the system arrived at its detection of stance. It aims to provide transparency by revealing the underlying pivotal arguments extracted from the article, allowing users to trace the system's decision-making process. (ii) **Insight into Author's Arguments**: The system's explanation may summarize and present the key arguments made by the author in the article. This helps users gain a better understanding of the author's perspective, the reasoning behind their assertions, the evidence they provide to support those assertions along with the credibility of the claim and form their own judgments. (iv) **Validation and Bug Identification**: In the development of the system, the explanation can help identify potential errors or biases. Users can scrutinize the extracted pivotal arguments and evaluate whether they accurately represent the main content of the article. This feedback can be used to improve the system's performance and ensure its reliability. Overall, the purpose of the explanation is to enhance user understanding, facilitate critical thinking, and provide valuable insights into the claim-article relationship and the reasoning behind the system's predictions.

In sum, our main contributions are the following.

- Unlike prior work on stance explanations, which confine themselves to snippets or words as opposed to well-formed sentences, we are the first to build a stance tree to combine evidence at different granularities and provide meaningful unsupervised stance explanations in the form of arguments justifying stance in a claim-article setting.
- Our proposed stance explanation architecture serves as a versatile and adaptable module that can seamlessly integrate with any stance detection

approach. Empirical experiments demonstrate that our explanation-based approach consistently achieves superior or on-par performance across a range of stance detection datasets, even when compared against robust and well-established baselines.

- We have crawled and prepared two datasets (Createdebate and Room For Debate) for Stance Detection in the claim-article setting. Between them, these datasets contain approximately 66,000 samples.
- Human studies show that our unsupervised technique of generating *stance explanations* outperforms the SOTA extractive summarization method in terms of informativeness, non-redundancy, coverage, and overall quality.
- Our pruning method yields a significant 16% improvement in Macro $F_1$-score when compared to the non-pruned representation. Additionally, it offers the added benefit of reducing the explanation length by an impressive 34%, thereby saving valuable time for the users.

## 2. Related Work

### 2.1 Stance Detection

Feature-based methods for SD have gained much popularity in online debate platforms (Anand et al. 2011; Somasundaran and Wiebe 2009, 2010; Krejzl, Hourová, and Steinberger 2017; Dey, Shrivastava, and Kaushik 2017). These models heavily rely on lexical features like $n$-grams, syntactic dependencies, and so forth. Though some of these works rely on sentiment analysis, they do not consider the importance of objectivity analysis from the perspective of SD. Somasundaran and Wiebe (2010) focused on identifying argumentative words, but this kind of lexical ontology severely hurts the model's performance when users directly present facts instead of rephrasing the facts in an argumentative manner.

Deep learning–based techniques have been extensively used for SD, especially for the target-tweet setting, since the introduction of SemEval-2016 task-6 (Mohammad et al. 2016). Earlier works on the target-tweet setting mostly relied on recurrent and convolutional neural networks, coupled with attention mechanisms (Rakholia and Bhargava; Ruder et al. 2018; Wei et al. 2016; Augenstein et al. 2016; Sun et al. 2018), whereas recent works (Jain, Doshi, and Kurup 2020; Gunhal et al. 2022) use transformer (Vaswani et al. 2017)-based architectures. Glandt et al. (2021) utilized self-training and knowledge distillation to leverage unlabeled tweets for COVID-19 tweet SD. Samih and Darwish (2021) focused on user-level stance detection by representing tweets using contextualized embeddings, which capture latent meanings of words in context. Clark et al. (2021) used knowledge graphs to find pathways connecting entities mentioned in a tweet to SD targets and later injected this knowledge in a RoBERTa model.

For dealing with the limited number of topics and to amplify the usefulness of SD models in unknown real-world scenarios, the zero-shot and few-shot SD setting has drawn a lot of attention (Allaway, Srikanth, and McKeown 2021; Luo et al. 2022), especially with contrastive learning (Liang et al. 2022; Liu et al. 2022). To enforce target dependency, Yuan et al. (2022) proposed two simplifying stance reasoning subtasks, determining whether the target is present in the text and whether the tweet has a stance towards the target. A second line of work focuses on the use of external signals

as contextual information for SD—for example, exploiting a user's interactions with her communities in social media (Del Tredici et al. 2019), identifying the required background knowledge for SD using Wikipedia (He, Mokhberian, and Lerman 2022), and incorporating stock market signals for Twitter SD (Conforti et al. 2022). All these works focus on determining the stance with respect to (w.r.t.) a topic (or a target), which is significantly different from determining the stance of an article w.r.t. a claim, owing to the semantic nuances present in a claim, unlike a topic.

SD in the claim-article setting attained popularity since the launch of the Fake News Challenge competition. The top-3 winning systems in this competition were Talos (Sean Baird and Pan 2017), Athene (Hanselowski et al. 2017), and UCLMR (Riedel et al. 2017), in this order. Talos applied a one-dimensional convolution neural networks (CNN) on the headline and body text, represented at the word level using Google News pretrained vectors. The output of this CNN is then sent to a multilayer perceptron with 4-class output: *agree*, *disagree*, *discuss*, and *unrelated*, and trained end-to-end. Zhang et al. (2019) addressed the problem by proposing a hierarchical representation of the classes, which combines agree, disagree, and discuss in a new related class. A two-layer neural network is used to learn from this hierarchical representation of classes and a weighted accuracy of 88.15% is obtained by their proposal. Furthermore, Dulhanty et al. (2019) constructed a stance detection model by performing transfer learning on a RoBERTa deep bidirectional transformer language model by taking advantage of bidirectional cross-attention between claim-article pairs via pair encoding with self-attention. They reported a weighted accuracy of 90.01%. Sepúlveda-Torres et al. (2021) attained the state-of-the-art performance by utilizing the automatically extracted summaries of articles instead of the full text along with features like overlap similarity, cosine similarity, and so on. Multi-task and multi-dataset learning have also been applied in the context of SD. Fang et al. (2019) showed that it is possible to improve SD performance by training on tasks like sentiment analysis, question answering, and textual entailment. Schiller, Daxenberger, and Gurevych (2021) leveraged the MT-DNN framework (Liu et al. 2019a) to simultaneously train on 10 SD datasets and attained SOTA results on most of the datasets. Unlike these works which focus on determining a *single* overall prediction given a claim and an article, we synthesize arguments from the article and make predictions for each of them. Later, we utilize these predictions in a hierarchical fashion to determine the final stance, thus allowing us to do fine-grained stance analysis.

Unsupervised SD approaches like Kobbe, Hulpuş, and Stuckenschmidt (2020) and Ghosh et al. (2018) make use of a lexicon and syntactic rules, which do not perform well when the articles contain objective reasoning, sarcasm, and negation. Pick et al. (2022) build an Interaction graph of the debate participants and utilize debate metadata such as the number of quotes and replies, making it inapplicable to our setting: There is no conversation present in a claim-article setting for SD.

Recently, SD has gained considerable attention for languages besides English, such as Spanish (Respall and Derczynski 2017), Italian (Cignarella et al. 2020), and Chinese (Xu et al. 2016; Agerri et al. 2021). In our work, we focus on SD in English and rely on tools that provide discourse parsing and segmentation that can vary significantly depending on the choice of language.

### 2.2 Stance Explanation

There are only a handful of studies that provide an explanation for their predicted stance. Du et al. (2017) and Li and Caragea (2019) used attention weights to interpret token contribution and treat the token with the highest weight as an explanation. Popat

et al. (2019) tokenized their input into phrases and sequentially feed them to their proposed BERT variant while measuring the change in stance prediction probability. Mohtarami et al. (2018) used an inference module to extract evidence from their input and ranked them based on their similarity with the claim. Compared with our work, *these systems cannot produce explanations beyond phrase level* and tend to *produce explanations that are incomprehensible* and only semantically related to the claim without expressing an actual stance. Conforti et al. (2020) focused on supervised evidence retrieval for SD in the claim-article setting by building and annotating a dataset from the financial domain. Sentences in articles are marked by experts if they serve as explanations for the chosen stance w.r.t. the claim. Jayaram and Allaway (2021) separated a small portion of the VAST dataset (Allaway and McKeown 2020) and asked annotators to determine the $k$ most important words in each article. Later, they used these human rationales as attribution prior to provide faithful explanations, that is, they added a loss function to minimize the difference between attribution prior scores and attention scores. We, on the other hand, focus on *unsupervised* stance explanations by revealing the argumentative structure of the entire article and do not restrict ourselves to any particular domain, thus avoiding the cost of manual annotation.

## 3. New Datasets for Claim-Article SD

The most commonly used datasets for SD have been the SemEval-2016 Task-6 (Mohammad et al. 2016) and the Fake News Challenge (FNC) Stage-1 datasets (Hanselowski et al. [2018]). The former consists of target-tweet pairs and is unsuitable for our claim-article setting. The FNC dataset is dominated (approx. 75%) by "unrelated" and "discuss" cases (i.e., no stance cases). To overcome the lack of a large and balanced dataset, we crawled a large number of claim-article pairs from a prominent online debate platform, Createdebate.[1] We refer to the dataset as Createdebate (CD). We also create a small but high-quality dataset where articles are collected from the *New York Times*'s opinion pages (Room For Debate), which we refer to as the RFD dataset. Below, we discuss the dataset retrieval process along with the annotation process for both datasets. We have released both CD and RFD for future research.[2]

### 3.1 Createdebate Dataset Retrieval

Createdebate is a popular online debate platform where users write their opinions on a variety of claims. To create a new dataset, we collected all the for/against debates from the Createdebate Web site, which focuses on pro and con stance labels. Each claim under the for/against category is accompanied by two predetermined sides. Users express their arguments in the form of a debate post and choose one of the sides which can potentially act as the chosen stance label towards the claim. Besides that, users can also reply to another user's post in the form of support or rebuttal. In this work, we do not focus on the analysis of hierarchical conversation structure; hence we discard all posts that are in the form of replies.

A previous version of the Createdebate dataset was created in Hasan and Ng (2014), which only focused on four popular domains, Abortion, Gay Rights, Obama,

1 https://www.createdebate.com/.

2 https://drive.google.com/drive/folders/1GWjMEcM1gvl9Kk4LCzGEr1-N4XgMc3Jo?usp=sharing.

and Marijuana. In our case, we crawled debates from all domains, thus building an SD dataset with a highly diverse set of more than 7,000 unique claims.

### 3.2 Createdebate Dataset Annotation Guideline

We identified two challenges with the crawled data:

1. The claim does not always express an opinion.
2. The collected debate sides are not always "pro" or "con."

A sample (claim, article, debate side) triple from the Createdebate Web site is shown below:

> *Claim*: Capitalism vs Communism
>
> *Article*: Capitalism is better. Communism enslaves each individual to the community. Capitalism allows each individual to be free and independent. Communism never makes progress because individuals are not given a just reward for improving things, while capitalism constantly improves. Communism lacks the economic freedom of choice to the fullest extent possible, in Capitalism people are free to choose what they want. Communism gives the lazy as much as the hard working, while capitalism allows each individual to sell their labor, creating a result where the hardest working are rewarded in accordance to how hard they work.
>
> *Chosen Debate Side*: Capitalism

Here, the claim "Capitalism vs Communism" does not express a clear stance toward either side. Also, the chosen debate side is not "pro" or "con." An ideal conversion for the claim above can be "Capitalism is a better economic system than Communism" because now it expresses a clear stance. Based on the modified claim and chosen side, the stance label can be inferred to be "pro." However, the conversion step is non-trivial. To see this, consider another example.

> *Claim*: Should the US intervene in Pakistan?
>
> *Debate side 1*: Pakistan would benefit
>
> *Debate side 2*: Not the business of the US

Here, while the claim expresses an opinion, it is hard to build a system that can automatically determine that any article that chooses side 1 is expressing a "pro" stance w.r.t. the claim, due to the semantic complexity. To better fit the collected data in our desired claim-article setting, we undertook a conversion step. We showed the crawled claims and debate sides to seven graduate students and asked them to

1. modify the claim so that it clearly expresses an opinion and
2. infer the stance label (either "pro" or "con") from the modified claim and the given debate sides.

We also showed 5 examples of correct and incorrect conversions to each student. The conversion took about two weeks and each student worked on approximately 1,000–1,500 claims. The payment was $8 per hour for each student.

After the conversion step, we had 77,770 articles and 8,561 claims. Since the conversion can be subjective, we performed an Amazon Mechanical Turk (AMT) experiment and asked turkers whether the conversion is acceptable. We first randomly selected 172 turkers whose number of HITs[3] approved is more than 10,000 (this is the maximum value that can be selected in AMT), and the HIT approval rate is more than 97%. We showed 50 conversion pairs to each turker and asked them to verify whether the conversion is correct or not. Each pair consists of a claim before and after conversion. The turkers were instructed to check two criteria: (1) whether, after conversion, the claim clearly expresses an opinion or not, and (2) whether the labels after conversion are coherent with the labels before conversion. If a conversion was decided to be unacceptable by a turker, we rejected it. Each turker was paid $15 per hour. After the conversion is complete, we eliminate any article with less than 3 words and replace URLs with the token "URL." The turkers rejected about 17% of the claims, yielding 64,654 articles with 7,135 claims. It is split into training (80%), development (10%), and test sets (10%).

### 3.3 Room For Debate Dataset Retrieval

The Room For Debate dataset is crawled from the *New York Times*'s opinion section where knowledgeable contributors discuss news events and other timely issues. Each claim is followed by a series of articles written independently by many authors. All claim-article pairs were crawled from the period of January 2009 to February 2017. An example of a collected claim-article pair follows.

> *Claim*: Big Data Is Spreading Inequality.
>
> *Article*: The White House report that sounded an alarm about the misuse of big data in housing, credit, employment and education also highlighted its ability to identify health risks at early stages in communities, create efficiencies in energy distribution and uncover fraud through predictive analyses. And while the Federal Trade Commission and others have pointed to the potential for discrimination in big data, it is important to remember that it can also advance the interests of minorities and actually fight discrimination. . . . In New York, there is a new campaign to collect coordinated data to improve health and human services for lesbian, gay, bisexual and transgender individuals, after a 2011 Institute of Medicine report showed that a "scarcity of research yields an incomplete picture of the L.G.B.T. health status and needs." Concerns about potential misuse of big data are fair and do deserve attention, but the newfound flood of information also represents an advance in the use of technology that has real potential for bettering society.

### 3.4 Room For Debate Dataset Annotation

Unlike Createdebate, the collected claim-article pairs are not annotated with a stance label on the RFD Web site. Initially, two annotators worked on assigning stance labels

3 A Human Intelligence Task, or HIT, is a question that needs an answer. A HIT represents a single, self-contained, virtual task that a worker can work on, submit an answer to, and collect a reward for completing.

**Table 1**
Properties of CD and RFD datasets; *top row*: Most popular topics in both datasets extracted using BERTTopic; *second and third rows*: Average absolute sentiment score and subjectivity score, computed using TextBlob; scores are in the range of [−1, 1] and [0, 1], respectively. A subjectivity score of 0.0 is very objective and 1.0 is very subjective. *bottom row*: Average Flesh Reading Ease (FRE) Score; FRE score in the range of 80.0–90.0 requires a school level of 6th grade to comprehend the text (easy to read), whereas a score between 50.0–60.0 demands a 10th to 12th-grade level (fairly hard to read).

| Properties | CD | RFD |
|---|---|---|
| Most Popular Topics | guns, Obama, abortion, god, animals, trump, racism, pets, parents, atheist | companies, college, nuclear, parents, politicians |
| Avg. Sentiment Polarity Score | 0.19 | 0.11 |
| Avg. Subjectivity Score | 0.47 | 0.45 |
| Avg. Flesch Reading Ease Score | 80.78 | 52.18 |

to each claim-article pair independently. They considered three possible stance labels:

1. "pro": if the article's overall stance is supporting the claim.
2. "con": if the article's overall stance is contradictory towards the claim.
3. "balanced": if the article presents an approximately even mixture of supporting and refuting arguments.

The inter-annotator agreement, expressed in terms of Cohen's Kappa score (Carletta 1996) was 82.85, showing a high agreement between the two annotators. In case of disagreements, a third annotator was asked to determine the stance and a majority voting was done in the end to assign the final stance.

### 3.5 Comparison between CD and RFD Datasets

In Table 1, we compare a set of syntactic and semantic properties between our two datasets, CD and RFD. First, we extract the most important topics from the claims, using BERTopic (Grootendorst 2022) for both datasets. Based on the identified topics, CD and RFD seem to have a very low overlap (only the topic "politics" is common among them). Next, We utilize the TextBlob (Loria 2018) tool on the articles of our datasets and report the average absolute sentiment polarity and subjectivity scores. According to Table 1, compared with CD, articles in RFD express less sentiment and more objectivity in general. We also aim to determine how readable (difficulty of reading) the articles are in CD and RFD and what type of reader can fully understand them. We compute the well-known readability index, Flesch Reading Ease (FRE) (Yeung, Goto, and Leung 2018; Spadaro, Robinson, and Smith 1980) for each article and present the average. Higher scores indicate material that is easier to read, lower numbers mark material harder to read. We hypothesize that because articles in RFD are written by experts who often utilize facts and statistics to convey their arguments, it is more likely they have a lower FRE score than CD. Table 1, *bottom row* verifies our hypothesis.[4]

4 https://simple.wikipedia.org/wiki/Flesch_Reading_Ease.

**Table 2**
Comparison of CD and RFD w.r.t. three other datasets (FNC, IAC, ARC) regarding total samples, class distribution, total unique claims, and median article length in terms of sentences and words.

| Dataset | Total Samples | Classes | # Unique Claims | Article Length (# sentences) | Article Length (# words) |
|---|---|---|---|---|---|
| CD | 64,654 | pro (49%), con (51%) | 7,135 | 2 | 39 |
| FNC | 75,385 | unrelated (73%), discuss (18%), agree (7%), disagree (2%) | 2,532 | 13 | 322 |
| RFD | 764 | pro (44%), con (47%), balanced (9%) | 215 | 17 | 416 |
| ARC | 17,792 | unrelated (75%), disagree (10%), agree (9%), discuss (6%) | 186 | 5 | 101 |
| IAC | 5,605 | pro (56%), anti (34%), other (10%) | 10 | 19 | 326 |

### 3.6 Comparison Against Other Datasets

In Table 2, we compare various statistical properties of our created datasets w.r.t. three other well-known datasets for claim-article SD: FNC,[5] ARC (Hanselowski et al. 2018), and IAC (Walker et al. 2012). Only the FNC dataset has a sample size larger than CD, although, CD by far dominates FNC in terms of the ratio of "pro" and "con" samples.[6] In terms of the number of unique claims, the CD dataset is by far the largest, almost 64% higher than its nearest competitor, FNC. On the other hand, the RFD dataset, despite its small size, has the second-highest article median length among the five datasets.

## 4. System Overview

The input to our system is a claim and an article. To build a stance tree, we first concatenate the claim and the article and then build a DPT. A DPT has EDUs as its leaves and its internal nodes express the combination of a collection of EDUs. To further reveal the argumentative structure of the article, we build a DDT from the DPT. A DDT only has EDUs (or only sentences) at every node. We use a stance predictor model to make a prediction for each (claim, EDU) or (claim, sentence) pair. However, not every EDU (or sentence) bears a stance toward the claim. We thus use pruning strategies to remove unnecessary elements from the DDT. We also implement (dis)agreement handling in the DDT. Once we have the stance predictions for every node in the DDT, we aggregate them using DST to reach the final prediction. In Figure 1, we only showed the input and output of our system, whereas in Figure 2, we illustrate each stage of our pipeline. In the following sections, we describe the various steps in detail.

### 4.1 Discourse Parse Tree

For each article, a DPT is generated based on RST. While most studies in the literature utilize a DPT at the EDU level ($DPT_e$), we also build a DPT at the sentence level ($DPT_s$) using the following rules:

- If an internal node $e_i$–$e_j$ in the $DPT_e$ contains EDUs from two different sentences $s_i$ and $s_j$, we add a node $s_i$–$s_j$ in the $DPT_s$.

5 http://www.fakenewschallenge.org/.
6 "pro" and "con" refers to the same stance as "agree" and "disagree"/"anti," respectively.

- If an internal node $e_i$–$e_j$ in the $DPT_e$ contains the start and end EDUs of a sentence $s_i$, we add a node $s_i$ in $DPT_s$.
- If a sentence $s_i$ has only one EDU $e_i$, then we simply add a node $s_i$ in the $DPT_s$.

Nodes in $DPT_e$ not belonging to the above criteria are ignored. Once we build $DPT_s$, we need to assign nuclearity to each of its nodes. If there is a one-to-one mapping between a node in $DPT_e$ and in $DPT_s$, we simply copy the nuclearity from $DPT_e$ to $DPT_s$. Otherwise, we count the number of nuclei present in each child sentence and use majority voting to designate one of the children as the nucleus. In case of a tie, we heuristically choose the left child to be the nucleus. Figure 2 (top row, Right) shows the DPT for the claim and article appearing in the top row, Left.

### 4.2 Discourse Dependency Tree

To better model the argumentative structure, we use a dependency tree, where every node corresponds directly to an EDU in $DPT_e$ (or a sentence in $DPT_s$). Generating the DDT from a DPT has been studied before (Li et al. 2014; Hirao et al. 2013). EDUs (or sentences) that are closer to the root of a DDT are more informative than those which appear at the leaf level. Moreover, the significantly smaller number of nodes in a DDT compared with a DPT allows us to quickly propagate results from the bottom to the root of the DDT and reach a global conclusion on the stance. Figure 2 (bottom row, Left) shows the DDT at the EDU level, constructed from the DPT shown in the top row (Right).

### 4.3 Stance Predictor Module

To predict the stance for every EDU (or sentence) in the DDT, we train a model $Model_{stance}$ (more details in Section 6). Later, we use it to make predictions for claim-EDU or claim-sentence pairs. Unlike existing works, which only focus on the prediction from a model, we store the prediction, $l_p$, along with the confidence, $l_c$, in that prediction for each node in the DDT. Later, we will use both of them to propagate evidence throughout the DDT.

### 4.4 Pruning Module

Because not every sentence plays an equal role in determining the stance, we add a pruning module to capture the most relevant arguments in the article. To determine the relevance of an EDU (or sentence), we experiment with a simple pruning strategy based on the similarity between claim and article EDUs (or sentences). We use a semantic similarity detection model $Model_{sim}$ to compute the cosine similarity between the claim and an EDU (or sentence), and treat EDUs (or sentences) whose similarity to the claim is below a set threshold as candidates for pruning. We prune the less relevant EDUs (or sentences) without violating the DDT construction algorithm while also preserving the nucleus-satellite relations. See Appendix A: Similarity Pruning Algorithm for the details of the algorithm. The pruning module takes the DDT created in subsection 4.2 and $SL$, a list of EDUs (or sentences) that are deemed important by $Model_{sim}$, and generates the pruned DDT. For example, in Figure 2 (bottom row, Middle), EDU 3 gets pruned because its similarity with the claim is lower than the threshold.

### 4.5 Detecting Agreement/Disagreement

In this module, we aim to detect (dis)agreement quantifiers and their respective scope for a proper understanding of the argument. One reason this may be essential is that breaking down an article into a DPT may easily separate the (dis)agreement quantifier from its scope. A quantifier itself may not express a stance but combining it with its scope has a higher potential of doing so. To detect (dis)agreement, first, we compile a set of agreement and disagreement quantifiers. We choose a subset of quantifiers from the Arguing Lexicon[7] corresponding to the categories "Assessments," "Doubt," "Emphasis," "Necessary," and "InyourShoes" and manually add some more (dis)agreement quantifiers by using synonymous words to the already selected quantifiers, leading to a total of 54 agreement and 43 disagreement quantifiers. We compute the similarity between every quantifier and every EDU using $Model_{sim}$; if the score is higher than a set threshold, we treat it as a (dis)agreement quantifier. After performing extensive experiments with the dependency parser, we found that the Clausal Complement (CComp) relation is extremely useful in detecting the scope of the quantifier. Once we detect the quantifier and its scope, we simply remove the quantifier if it is of type agreement. If it is a disagreement quantifier, we also flip the stance predicted by $Model_{stance}$ for the EDUs in its scope. In Figure 2 (bottom row, Middle), for example, EDU 5 gets pruned because it is a disagreement quantifier (its scope is EDU 6). However, EDU 6 is marked as 6* to realize that, prior to aggregation, its predicted stance will be flipped.

In case we are unable to determine the scope or the scope is empty, we treat the (dis)agreement quantifier as a standalone (contradictory) supporting evidence towards the claim. We determine whether the stance label for this should be pro or con depending on whether it was an agreement or disagreement quantifier.

### 4.6 Aggregation

Once we obtain all the predictions for each sentence or EDU in the pruned DDT, we need a way to propagate the predictions from the leaf nodes to the root of the DDT to determine the final stance label. Instead of heuristically resorting to averaging or taking a majority vote, we use Dempster's Rule of Combination (Sentz et al. 2002; Voorbraak 1991; Lefevre, Colot, and Vannoorenberghe 2002; Murphy 2000) to combine evidence. Two bodies of evidence can be combined with Dempster's rule as follows:

$$m_{\{1,2\}}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K} \tag{1}$$

where $K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$.

Here, for a set of evidence S, $m_S(A)$ denotes the (combined) support of evidence $S$ for $A$ and $K$ is a normalizing constant (Sentz et al. 2002). A, B, and C correspond to stance labels (in our case, either pro or con). We write $m_1$ instead of $m_{\{1\}}$ for simplicity. For every node, we treat the prediction, $l_p$ as "evidence" and the confidence score, $l_c$ as the "weight" of the evidence.[8] After applying the rule, the newly computed prediction and confidence of a node are $g_p$ and $g_c$. We apply the rule in a bottom-up manner, that is, based on its $l_p$ and $l_c$ and its children's $g_p$ and $g_c$.[9] In Figure 2 (bottom row, Right), we

7 https://mpqa.cs.pitt.edu/lexicons/arg_lexicon/.

8 $l_p$ and $l_c$ are computed using $Model_{stance}$ as described in subsection 4.3.

9 For leaf nodes, local and global predictions (and confidences) are the same.

show the stance predictor's prediction and confidence with Dempster's rule. We explain the application of the DST rule in the stance tree for node 7. Here, $m_7(pro) = l_c(7) = 0.79$ and $m_6(pro) = g_c(6) = 0.84$. Using DST, we combine them to compute $m_{\{7,6\}}(pro) = g_c(7) = 0.95$.

## 5. Human Evaluation on Explanations

In this study, we define **stance explanation** as the extraction of crucial arguments from an article to clarify the author's position towards a claim. This section presents a human evaluation study aimed at examining the quality of the explanations generated by our approach. In the context of stance explanation, we draw a parallel to the notion of "precision" in machine learning, focusing on the relevance and accuracy of the extracted sentences in conveying the author's stance. Precision assumes a pivotal role in the extraction of sentences as stance explanations, as it evaluates how well the selected sentences align with the author's actual stance. High precision signifies that the extracted sentences effectively represent and support the author's intended stance, resulting in a compelling explanation. It also ensures that the extracted sentences are highly relevant and accurately capture the author's intended perspective.

Following previous work (e.g., Mao et al. 2020; Wu and Hu 2018; Narayan, Cohen, and Lapata 2018; Luo et al. 2019), we intend to determine the quality of the explanations on the following four criteria: Informativeness, Non-Redundancy, Coverage, and Overall Quality, defined below:

- *Informativeness*: How informative is the explanation for justifying the stance label w.r.t. the claim?
- *Non-Redundancy*: Does the explanation contain redundant arguments?
- *Coverage*: To what extent does the explanation provide multiple unique perspectives for justifying the stance label?
- *Overall quality*: Overall, how satisfactory is the explanation for justifying the stance?

Regarding Coverage, we treat coverage as analogous to recall. Within the scope of our study, an article may contain numerous relevant arguments supporting the author's stance toward a claim. We measure coverage to assess a system's ability to provide as many relevant arguments as possible within a given budget.

### 5.1 Choice of Baselines

The natural candidates for baselines for experimental comparison would be stance explanation systems such as Conforti et al. (2020) and Jayaram and Allaway (2021). Instead, we choose *T*5 and *MS*, which are summarization systems, as the baseline explanation systems. We next justify this choice. Conforti et al. (2020) collected articles about four recent mergers involving US companies and asked experts to select the text snippets or sentences from the article that were determinant for them to classify their stance. We believe that a supervised method trained on a particular domain (for example, finance) will perform poorly when asked to generate explanations for SD in significantly different domains such as education, healthcare, politics, and so forth, due

to the challenges of transfer learning. While our approach and that of Conforti et al. (2020) to stance explanation fall on the opposite sides of a spectrum (unsupervised vs supervised), the approach of Jayaram and Allaway (2021) falls somewhere in the middle of that spectrum. Indeed, unlike Conforti et al. (2020), they only annotate a small portion of the training examples with stance explanation and perform few-shot and zero-shot experiments. However, Jayaram and Allaway (2021) asked annotators to mark the most important $k$ words in each article, where importance depends on whether masking that word would enhance the difficulty of determining the stance. This kind of span-marking approach is significantly different from ours because we aim to provide an explanation by revealing the argumentative structure of the article.

We briefly present an example of how our approach would compare with supervised approaches. In Figure 3, we compare explanations generated from our system (Right) and Jayaram and Allaway (2021) (text span marked in red on the left). For the claim, "Home birth is a safe choice," Jayaram and Allaway (2021) pick phrases such as "choice," "best managed," "safe option," and so forth, which are at best semantically related to the claim. On the other hand, our proposed stance tree manages to include

**Claim**: Home Birth Is a Safe Choice. (stance label: pro)

**Article**:
s1: The controversy about whether or not women should give birth at home is moot.
s2: In the US today, home birth is a choice women are already making in increasing numbers.
s3: The real question we need to answer, ... is: What conditions make home birth a safe option?
s4: Studies of large numbers of women who planned home births show which pregnancy- specific conditions are best managed in a hospital.
...
s8: But, the question of safety also implicitly asks ... In fact, the majority of labor and birth complications are non-emergencies.
...
s10: Life-threatening emergencies that are not predictable are extremely rare — ... equal rates.
...
s11: In Britain, the national health service actually advises healthy women that it is safer to have children at home, ... when the pregnancy is low-risk.
s12: Still, in the U.S., the home birth process could be improved ... when complications appear after labor starts or the infant is born.

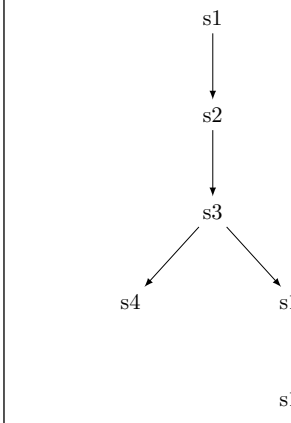



**Figure 3**
Stance explanation comparison between Jayaram and Allaway (2021) and our approach. *Left*: a sample claim, an article, and the stance label; sentence indices marked in blue correspond to sentences that belong to the explanation generated from our proposed system; text spans marked in red belong to the explanation generated from Jayaram and Allaway (2021). *Right*: Stance tree based explanation, generated from our system.

the most important arguments such as $s1$ and $s2$ as part of the explanation. Even the "flat" version of our explanation, presented without a stance tree, is clearly far more informative than the explanation of Jayaram and Allaway (2021).

Furthermore, there is a significant difference between the existing body of work on generating stance explanations and our approach. Most of the existing work (Du et al. 2017; Li and Caragea 2019; Popat et al. 2019) follows a pipeline-like approach, where any change in the stance detection module will most likely require a major design change in the stance explanation module. However, our approach utilizes discourse parsing to construct the stance tree, which remains independent of the evidence aggregation scheme or stance predictor model. This modular design provides flexibility and allows for easier adaptation to different components without requiring a complete overhaul. Due to these shortcomings of supervised approaches, we compare our approach against powerful summarization systems that are known to generate condensed outputs that convey important information contained in an article. Extractive summarization aims to generate concise and coherent summaries by selecting the most important sentences from the original content. Through this comparison, we can effectively evaluate the efficacy and quality of our stance explanation system in providing meaningful and coherent explanations. By contrasting it with summarization methods, we can highlight the distinct advantages and contributions of our approach, showcasing how it goes beyond mere summarization to offer detailed, argumentative insights that shed light on the underlying stance of the article. Besides, we are interested in measuring the precision and recall-like qualities of our generated explanations. Extractive summarization also focuses on similar metrics (Graham 2015; Jia et al. 2020), making it a relevant and appropriate baseline for comparison. By assessing precision and recall, we can evaluate the effectiveness of our stance explanation system in accurately capturing pivotal arguments and providing comprehensive coverage.

To evaluate the quality of explanations generated by our system, we compare them against $T5$ (Raffel et al. 2019) and Matchsum (*MS*) (Zhong et al. 2020) via two large-scale human surveys. T5 is a text-to-text transformer-based language model which has shown outstanding performance in many NLP tasks. Matchsum, to our knowledge, is the SOTA extractive summarizer.[10] We refer to our system, eXplainable Stance Detection (*XSD*) without pruning as $XSD_{np}$ and with pruning as $XSD_p$.

### 5.2 Survey Designs

We design two separate surveys (hosted using AMT and Qualtrics platforms) to answer the following question about our system-generated explanations: *Does $XSD_{np}$ (and $XSD_p$) produce better explanations compared with other systems as we increase the article length?* In the first survey, we compare explanations generated by $XSD_{np}$, $T5$, and *MS*. As we believe that for many applications, a short explanation for the stance can be valuable, in the second survey we replaced $XSD_{np}$ with $XSD_p$, and compared it against $T5$ and *MS*. We perform two separate surveys because our goal is to evaluate the quality of the explanations with and without pruning. While designing our survey, we followed extractive summarization works like Narayan, Cohen, and Lapata (2018) and Mao et al. (2020) where authors have explicitly fixed the number of sentences determined from their proposed methods along with the baselines during experiments. We align the length of explanations generated from T5 or MS to be close to $XSD_{np}$ (in the first survey)

10 http://nlpprogress.com/english/summarization.html.

and $XSD_p$ (in the second survey). This allows us to better analyze and understand the impact of pruning in the generated explanations and eliminate any potential bias towards the explanation length.

Because our proposed stance tree can potentially preserve all the sentences in an article, we decided to show a shortened version of the tree in the survey by selecting the nodes closer to the root and with high confidence. For any sentence, $s$ in the stance tree that satisfies either of the following two criteria is shown in the survey: 1. $pos(s) \leq d_t$, 2. $pos(s) > d_t$ and $g_c(s) \geq c_t$, where $pos(s)$ denotes the depth of the sentence $s$ in the stance tree, $d_t$ and $c_t$ are depth and confidence threshold, respectively, and $g_c(s)$ refers to the global confidence in the prediction of the sentence $s$ after using DST (subsection 4.6). To ensure that participants can finish the survey in a reasonable time, we decided to show maximum 10 sentences in each stance justification sample, and we achieved that by setting $d_t$ and $c_t$ to 4 and 0.95, respectively.

As for explanations generated by $T5$, we limit their length to $w_c \pm k$, where $w_c$ is the number of words in the explanations generated by $XSD_{np}$ (or $XSD_p$ in the second survey), and we set $k$ to 10. For *MS*, following the guidelines from the authors, we limit the number of candidate combinations to 500 due to our system hardware's memory limit. We also set the number of sentences in the candidate summary to be the same as the number of sentences in $XSD_{np}$ (or $XSD_p$ in the second survey).

As we hypothesize that the length of an article is a key factor, we picked the RFD dataset as it offers a wide range in length. We separated RFD articles that have been correctly predicted by *both* $XSD_{np}$ and $XSD_p$, and split them based on article length into three bins—small, medium and large. The bins are created by following an approximate equi-depth histogram on article length: articles with $< 16$ sentences go into the small bin; those with 16–20 sentences into the medium bin; and those with $> 21$ sentences into the large bin. We randomly picked 30 examples from each bin.

### 5.3 Survey Setup

Recall that our human evaluation consists of two surveys. Prior to running both surveys, we hosted a qualification test on AMT to select the best candidates for the survey and we only allowed candidates whose number of HITs approved is more than 10,000 and the HIT approval rate is more than 97%. We asked five questions regarding stance detection and justifications and only turkers who scored at least 4 out of 5 were invited to participate in the survey.

After the qualification test, 80 turkers met our criteria for the first survey and we asked them to answer 10 questions in 1 hour. In each question, we showed them a claim, the stance label, and the explanations generated from the aforementioned systems. Our goal was to have at least five responses per question. In the end, 72 out of the 80 turkers participated in the survey and we had on average 6.1 responses per question.

For the second survey, we followed the same procedure as we did in the first survey. However, during the qualification test, we excluded all the turkers who participated in the first survey. After the qualification test, 66 turkers met our criteria and we asked them to answer 10 questions in 1 hour. In the end, 52 out of the 66 turkers participated in the survey and we had on average 5.88 responses per question. We set the time limit for the qualification tests and the survey to 12 minutes and 1 hour, respectively, and turkers were paid at a $15 hourly rate.

For both surveys, we asked the turkers to score the explanations on all four evaluation criteria, described at the beginning of Section 5 on a scale of 1 to 5 (1 being the worst, 5 being the best). More details on the survey guidelines can be found in

**Table 3**
Example of a question shown in the second survey where we compare explanations generated by *T5* (Snippet 1), $XSD_p$ (Snippet 2), and *MS* (Snippet 3).

| |
|---|
| **Claim:** Foreign Language Classes should Be Mandatory in College.<br>Stance Label: **con** |
| **Snippet 1**<br>1. high school students look forward to college as a place where they are free to choose what they want to study.<br>2. learning thinking skills, or even communication, isn't unique to foreign language classes, says physicist edward mccartney, who studied in laos for 11 years, can make real fluency seem unattainable.<br>3. the level that students |
| **Snippet 2**<br>1. While learning a foreign language theoretically presents a valuable opportunity to communicate or think in a different way, a foreign language requirement is problematic. (con)<br>1.1. First, it assumes that the same process of learning new "languages" and ways of thinking can't be accessed by simply studying a different discipline; and second, that students will learn a language well enough to actually experience a culture. (con) |
| **Snippet 3**<br>1. High school students look forward to college as a place where they are free to choose what they want to study and what career path they want to take.<br>2. While learning a foreign language theoretically presents a valuable opportunity to communicate or think in a different way, a foreign language requirement is problematic. |

Appendix B: Survey Guidelines. In Table 3, we show an example that was shown to the turkers during the second survey. The example contains the claim, the stance label, and stance explanations generated from *T5*, $XSD_p$, and *MS* systems, respectively. An example question from the first survey is shown in Appendix C: Example Question for First Survey.

### 5.4 Explanation Presentation

When presenting explanations, we adopt different formats for Matchsum and T5 compared with our approach, which incorporates a hierarchical structure through discourse parsing. An illustrative example of how we present explanations from different systems is provided in Table C.1. For Matchsum and T5, we present explanations in a straightforward list format. In the case of our approach, which generates explanations with a hierarchical structure, we use a specific method to ensure the preservation of this hierarchy. Starting from the root of the stance tree, we utilize the Breadth-First Search strategy to select and add sentences that contribute to the formation of the explanation. The selection process continues until it meets the criteria described in subsection 5.2. To maintain the parent-child relationship between arguments in our system's explanations, we utilize <tab> indentation along with hierarchical numbering when presenting them to the user. These visually represent the parent-child relationship among arguments, reinforcing the understanding of their interconnections within the explanation. In Snippet 2 of Table C.1, which corresponds to the explanation from our system, we use <tab> indentation along with numbering (e.g., 1, 1.1, 1.1.1) to highlight the hierarchical relationships among the arguments.

### 5.5 Results Analysis

In Figure 4, top row, we show the results from the first survey, which compares the quality of generated explanations from $XSD_{np}$, T5, and MS for all three bins—Small (Sm), Medium (Md), and Large (Lr). We show the percentage of the times, each model is ranked 1st in each bin. T5 performs the worst by far. $XSD_{np}$ holds the first position on non-redundancy more often than other systems across all bins. On the other criteria, $XSD_{np}$ is dominated by MS in the small bin, while the performance gap narrows as we move from the smaller bins to larger bins. For example, the performance gap closes from 37% (Sm) to 25% (Md) on informativeness, 24% (Sm) to 0% (Lr) on coverage, and 56% (Sm) to 0% (Md) in overall quality. This strengthens the argument that our system manages to deliver high-quality explanations as the article length increases. In Figure 4, bottom row, we show the results from the second survey, comparing the quality of explanations generated from $XSD_p$, T5, and MS for all three bins. We can see that summaries generated from our pruned stance trees are consistently preferred by humans over explanations generated by other methods in all bins and across all 4 evaluation metrics, proving the benefits of pruning for long articles.

In Figure 5, we show the results from the second survey, comparing the quality of explanations generated from $XSD_p$, $T5$, and $MS$ for the large bin. For lack of space, we only include the results for the large bin in the main paper and include the results for small and medium bins in Appendix D: Survey Result for Small and Medium Bins. We can see that summaries generated from our pruned stance trees are consistently preferred by humans over explanations generated by other methods across all 4 evaluation metrics, proving the benefits of pruning for long articles. Specifically, compared with Matchsum, our system shows a 23.2%, 11.4%, 26.9%, and 27.5% improvement in terms

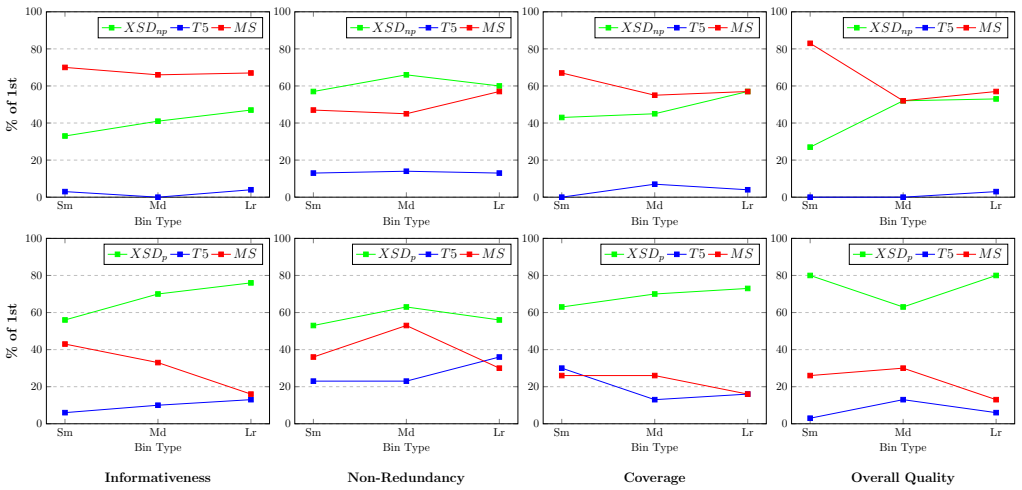



**Figure 4**
**Top row –** Comparison of stance explanations generated from our system without pruning ($XSD_{np}$), $T5$, and Matchsum ($MS$) across small (Sm), medium (Md), and large (Lr) bins. **Bottom row –** Comparison of stance explanations generated from our system with pruning ($XSD_p$), $T5$, and Matchsum ($MS$) across small (Sm), medium (Md), and large (Lr) bins. We plot the percentage of time each system was ranked 1st for each bin.

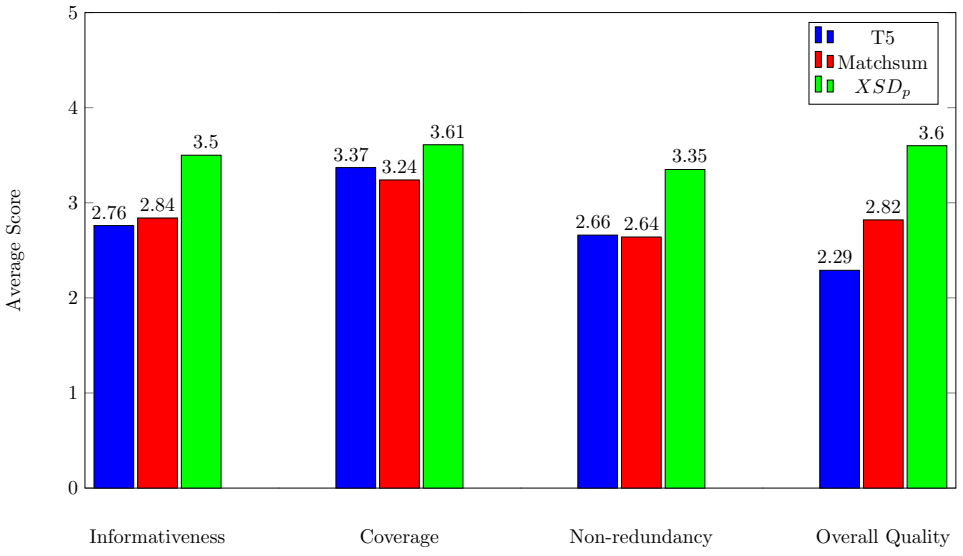



**Figure 5**
Comparison of the obtained average score (range is 1 to 5) for the stance explanations generated from our system with pruning ($XSD_p$), $T5$, and Matchsum ($MS$) across the four evaluation criteria in the large bin.

of informativeness, coverage, non-redundancy, and overall quality. We also witnessed similar traits in the small and medium bins as well.

### 5.6 Error Analysis

We perform a fine-grained error analysis of our system-generated explanations here. We divide the error analysis into two cases: case 1, where turkers found explanations generated from our system to be inferior compared to other systems, and case 2, where turkers deemed our explanations to be the best among the three systems, but they got an overall low score nevertheless. We show an example for each case in Figure 6 and Figure 7, respectively.

In Figure 6 (Right), we show the explanations generated from our proposed method, $XSD_p$ and $MS$ for the claim and article on the left. $XSD_p$ generated explanation misses some crucial arguments such as $s1$, $s3$, and $s10$. Further investigation shows that these arguments are absent from the stance tree because of the pruning strategy. Our pruning strategy largely relies on the similarity between the claim and an argument. This can fail for complex arguments. One possible way to overcome this is to rely on techniques from the argumentation mining domain and leverage external models that aim to identify the argument components in a document.

In Figure 7, we show the explanations generated from $XSD_p$ and $XSD_{np}$ for the claim and article on the left. This is an example where turkers preferred $XSd_p$'s explanation over other systems ($MS$ and $T5$) but were assigned a very low overall quality score nevertheless. The article shown on the left contains arguments that take on a "pro" ($s1$ until $s10$) as well as those that take on a "con" ($s10$ - $s24$) stance. Despite the fact that

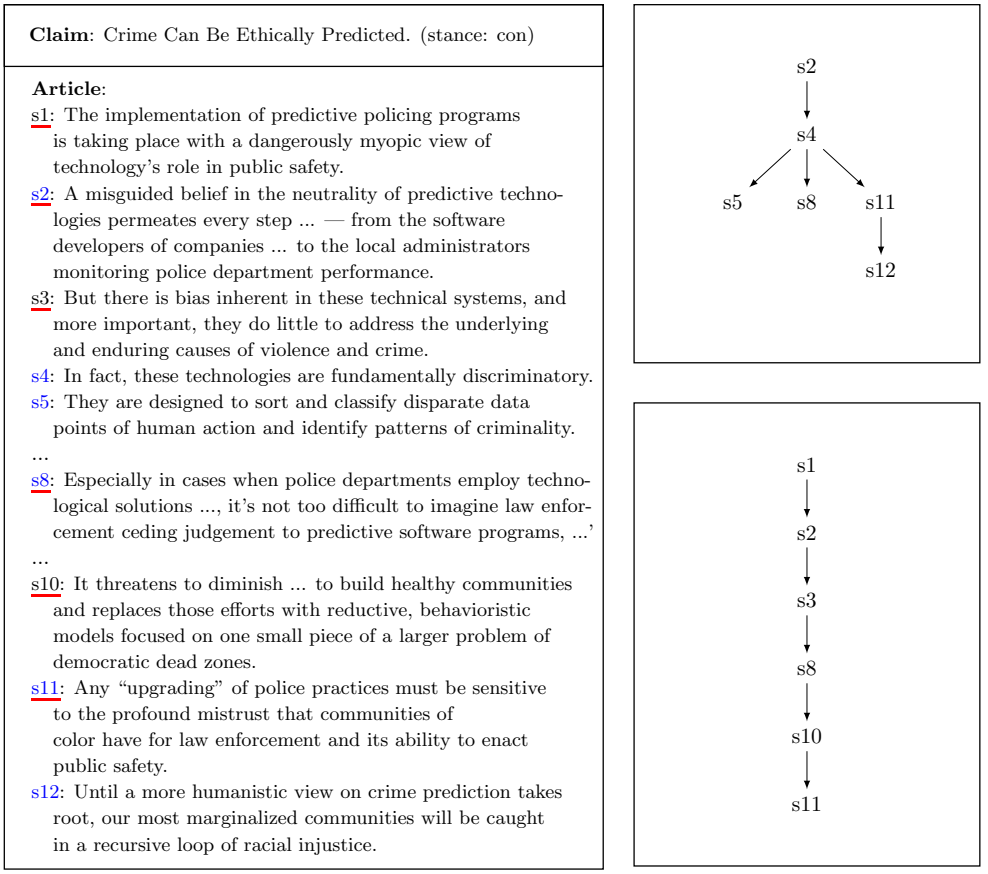



**Figure 6**
Error analysis, case 1: *MS*'s generated explanation is preferred by the turkers over $XSD_p$'s explanation in terms of overall quality. *Left*: A claim, article, and the stance label. Sentence indices marked in blue correspond to sentences that are part of $XSD_p$'s explanation. Sentence indices underlined in red correspond to sentences that are part of *MS*'s explanation. *Right-top*: Explanation generated from $XSD_p$; *Right-bottom*: Explanation generated from *MS*.

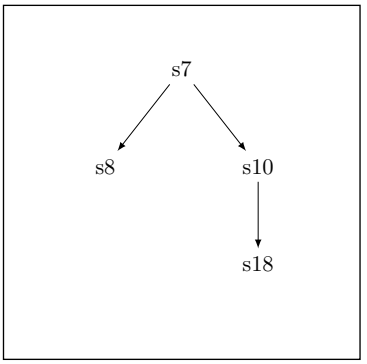
the "con" arguments drive the overall stance of the article w.r.t. the claim, none of those arguments are present in $XSD_p$'s generated explanation. We also show the explanation generated from $XSD_{np}$ (Figure 7, Right-bottom). Notice that the "con" arguments are absent from the non-pruned stance tree as well. Clearly, this has nothing to do with the pruning strategy but results from the imperfect nature of discourse trees. Ideally, the DDT for this example should have the "con" arguments closer to the root, which is not the case.

### 5.7 Comparison between T5 and GPT-3

In recent years, the GPT-3 (Generative Pretrained Transformer 3) model (Brown et al. 2020) has gained significant attraction owing to its extraordinary capability of producing human-like text. In addition, it has outperformed the SOTA models on tasks like

**Claim**: Anti-White Bias Is a Problem. (stance: con)

**Article**:
s1: When I got admitted to Yale and to Harvard, ...
I probably was taking up a white person's space.
s2: Sure, I was "qualified," but without diversity programs
..., African-Americans rarely end up in elite institutions.
...
s4: So, yes, when this black man got his, some random
white person lost out.
s5: Racism is not a zero-sum game, but opportunities
for limited resources are.
s6: This includes jobs that pay well and a quality
education.
s7: When white people perceive that they are
disadvantaged by a level playing field, they are correct.
s8: We've always known that anti-black bias is
purposeful; it turns out, though, ... purity.
...
s10: But, lest anyone worry, white folks, comparatively
speaking, are doing just fine.
...
s15: Given how much worse off African-Americans are
than whites, the most surprising thing is that many
blacks believe there has been so much racial progress.
s16: Based on the facts, their optimism is as
unfounded as the white's pessimism.
...
s18: But the president certainly understands that ...
anti-black bias is a historic relic.
...
s24: African-Americans still are owed big time.



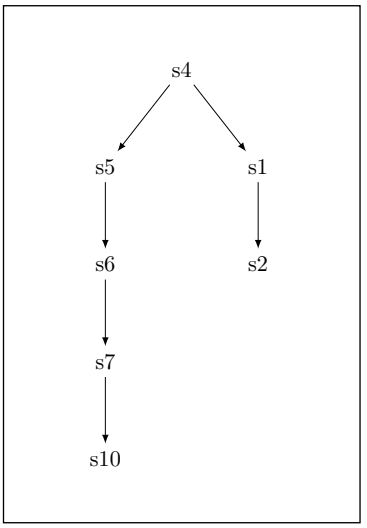




**Figure 7**
Error analysis, case 2: $XSD_p$'s explanation is preferred by the turkers compared to other systems; however, it gets a very low score in terms of overall quality. *Left*: A claim, article, and the stance label. Sentence indices marked in blue correspond to sentences that are part of $XSD_p$'s explanation. Sentence indices underlined in green correspond to sentences that are part of $XSD_{np}$'s explanation. Sentence indices marked in red correspond to sentences that are crucial in explaining the "con" stance of the article w.r.t. the claim but are not part of $XSD_p$ or $XSD_{np}$. *Right-top*: Explanation generated from $XSD_p$; *Right-bottom*: Explanation generated from $XSD_{np}$.

question answering, translation, and natural language inference, especially in the few-shot and the zero-shot settings. Given its superior performance over other language models, a natural question that arises is whether GPT-3 can generate better stance explanations compared to the T5 model.

To answer this question, we run an experiment where we leverage the zero-shot setting for both models and ask them to generate stance explanations. We set the prompt to be, "Determine the stance of the article w.r.t. the claim and explain why the article is taking that stance." followed by a set of claim-article pairs from the RFD

**Table 4**
*Left*: A set of claims and the stance labels. Similar to our Turk experiments, we do not show the articles. *Right*: Stance explanations generated from GPT-3 and T5.

| Claim (Stance Label) | Stance Explanations |
|---|---|
| Trump's infrastructure plan can work.<br>label: con | **GPT-3**: The article takes a negative stance towards Trump's infrastructure plan. The article argues that the plan is not the best way to address the problem and that it is not likely to meet the needs of the economy. |
| | **T5**: Trump is right to emphasize the need for more and better infrastructure, but his program is not the way to address the problem. there is much research showing the benefits of spending on traditional infrastructure such as roads and bridges. Ideally, a public investment agenda would carry over into areas like early childhood education, which we know provides huge benefits to the children directly affected and the economy over the longer term. |
| Voting should be a Constitutional Right.<br>label: pro | **GPT-3**: The article takes the stance that voting should be a constitutional right. The article argues that this would confirm our commitment to democracy. |
| | **T5**: Amercians without photo ID are disproportionately low-income, disabled, minority, young and older voters. voter ID laws, ending same-day registration and early-voting, and slowing the move to online voter registration are all popular ways to make it harder to vote, according to the Brennan Center for Justice. Voters in 15 states are finding it's much harder this year than it was in 2012... |
| All Immigrants Should Get ID Cards.<br>label: con | **GPT-3**: The article takes a negative stance with respect to the claim. |
| | **T5**: The mayor is willfully and explicitly trying to undermine federal immigration law. The purpose, as the mayor has made clear, is making it easier for illegal immigrants to open bank accounts, access city services, rent apartments, etc. All legal immigrants already have federally issued city identification cards that work just fine. he has said the card is not a reliable form of identification and should not be accepted. Mexico does not verify identity when issuing them, nor does it keep |

dataset. We report our findings in Table 4 for three such claim-article pairs. For the first example, explanations generated from GPT-3 and T5 are comparable. However, for the second example, the stance explanation generated from GPT-3 is much shorter compared with T5's (despite both being given the same summary length budget). In the third example, GPT-3 only determines the stance and completely ignores producing any kind of explanation. Due to the uncertainty of GPT-3 generated explanations, we decided to use T5 in our AMT experiments instead of GPT-3.

## 6. Stance Detection on 5 Datasets

In this study, our main emphasis is on the task of stance explanation. However, we also seek to demonstrate the compatibility of our proposed architecture with any stance detection model. The central question we aim to address is as follows: Can our model deliver high-quality explanations while achieving comparable performance in stance detection when integrated as a plug-in to any stance detection model? To investigate its adaptability, we evaluate its performance when coupled with various stance detection models. By integrating our approach as a plug-in, we assess its ability to generate high-quality explanations without compromising the overall performance of the underlying stance detection model.

In this section, we describe the experiments for evaluating the performance of our stance tree based stance detection approach on the five datasets (CD, RFD, ARC, FNC, and IAC) we introduced in Section 3. We show the test split along with the median

**Table 5**
Test Splits & Article Length Distributions.

| Dataset | Test Samples | Class Distribution | Article Length (#sentences) | Standard Deviation | Article Length (#words) |
|---|---|---|---|---|---|
| CD | 6,301 | pro (49%), con (51%) | 2 | 3.85 | 39 |
| ARC | 3,559 | pro (9.38%), con (10.45%), other (80.16%) | 5 | 2.75 | 101 |
| FNC | 25,413 | pro (7.49%), con (2.74%), other (89.77%) | 13 | 13.09 | 322 |
| RFD | 753 | pro (43.82%), con (47.14%), other (9.03%) | 17 | 5.61 | 416 |
| IAC | 924 | pro (42.75%), con (46.86%), other (10.39%) | 18 | 129.35 | 326 |

length of the articles for each dataset in Table 5. More details on the FNC, ARC, and IAC datasets including training and validation splits can be found in Schiller, Daxenberger, and Gurevych (2021). For discourse parsing, we use the segmenter introduced by Wang, Li, and Yang (2018)[11] and for the parser, we use the method introduced by Wang, Li, and Wang (2017).[12] To build the DDT, we follow Li et al. (2014).[13] We achieved the best Macro $F_1$ score on the development set when we chose 0.4 as the similarity threshold for the IAC and FNC datasets and 0.2 for the other datasets and 0.8 as the agreement threshold for all datasets. For $Model_{sim}$, we use the sentence transformers framework (Reimers and Gurevych 2020). We compare against the following baselines[14]:

- **MTDNN** (Schiller, Daxenberger, and Gurevych (2021))[15]: The proposed model learned from ten SD datasets of various domains in a single dataset learning setup (SDL) (i.e., training and testing on all datasets individually) and in an MDL setup (i.e., training on all ten datasets jointly). The authors used BERT (Devlin et al. 2018) and MT-DNN (Liu et al. 2019a) models; for each setup, we report the result for the best-performing architecture.
- **HSD** (Sepúlveda-Torres et al. (2021))[16]: The proposed model attained the best performance on the FNC dataset considering examples of all four stance classes. The authors created an extractive summary of the given article, then utilized a RoBERTa (Liu et al. 2019b) classification model to get a joint representation of the claim and the summarized article and infused it with various stance features.
- **MoLE** (Hardalov et al. (2021))[17]: The proposed framework combines domain-adaptation and label embeddings for learning heterogeneous target labels and obtains sizable performance gains over strong baselines, both (i) in- domain (i.e., for seen targets) and (ii) out-of-domain (i.e., for unseen targets).

11 https://github.com/PKU-TANGENT/NeuralEDUSeg.
12 https://github.com/yizhongw/StageDP.
13 For discourse parsing and segmentation, we used existing codes; for discourse dependency tree, we implemented Algorithms 1, 2, and 3 from Hayashi, Hirao, and Nagata (2016).
14 For the baselines, we used existing codes.
15 https://github.com/UKPLab/mdl-stance-robustness.
16 https://github.com/rsepulveda911112/Headline-Stance-Detection.
17 https://github.com/checkstep/mole-stance/tree/main.

Our primary objective is to provide stance explanations specifically for the "pro" and "con" classes. To ensure the compatibility of our proposed approach with stance detection methods that are not tailored to specific dataset variations, we implement the following steps. For each dataset, we consolidate any claim-article pairs that do not fall into the "pro" or "con" classes into a single category, which we refer to as "other." This consolidation allows us to handle the diverse variations among the datasets. We evaluate our approach in two stages. Initially, we train our base predictor model solely on the "pro" and "con" classes, and the results of this training are described in subsection 6.1. Our aim in this first stage is to demonstrate that our system is capable of generating high-quality explanations while maintaining performance comparable to the baseline models. Next, we leverage the trained model, which was designed for the two-class scenario, to make predictions for all three classes: "pro," "con," and "other." We achieve this by utilizing a simple heuristic-based rule (more details in subsection 6.4). The objective of our experiment in the second stage is to ensure the robustness and seamless compatibility of our approach with any system capable of performing stance predictions across more than two classes. By conducting these two stages of experimentation, we aim to showcase both the quality and versatility of our approach. It demonstrates that our system can deliver high quality explanations while achieving performance on par with baselines in the first stage, and it proves the adaptability of our approach to systems that can handle stance predictions across multiple classes in the second stage. To answer the research question introduced at the beginning of this section, we first utilize *MoLE* as the $Model_{stance}$ and report our findings in subsection 6.1. For any baseline $B$, we use the notation $B + XSD_p$ to refer to the case when $B$ is used as $Model_{stance}$ in our *stance tree* based approach with pruning. Also, the reported performances of all baselines are obtained by providing the entire claim-article pair as input to all the systems. We only show the performance at the sentence level here; for analysis on the EDU level as well as analysis on the aggregation policies used, see Appendix G: Result Analysis on EDU Level and Appendix F: Comparison of Aggregation Policies, respectively. We have published the code.[18] More details on the parameter settings and the baseline training are provided in Appendix E: Baseline Implementations.

### 6.1 Stance Detection Performance Comparison for Two Stance Classes

Based on the median article length, as indicated in Table 5, we classify IAC, RFD, and FNC as long article datasets, while ARC and CD are categorized as short article datasets. In Figure 8, we present the relative improvements and deteriorations in the performance of our approach compared to the *MoLE* framework. Notably, our approach (*stance tree* with *MoLE* as $Model_{stance}$) outperforms the *MoLE* framework on three out of the five datasets, with relatively minor performance degradation observed on the Createdebate dataset. These results provide compelling evidence that our approach consistently performs at least as well as, if not better than, the *MoLE* framework when used as $Model_{stance}$.

To gain further insights, we delve deeper into this phenomenon while using *MTDNN* and *HSD* as the chosen $Model_{stance}$ (Table 6). When using *MTDNN* as the stance predictor model, our approach demonstrates on-par performance on three out of the five datasets, exhibiting only marginal performance drops on the IAC and FNC datasets. On the other hand, with *HSD* as the stance predictor model, our approach exhibits

18 https://anonymous.4open.science/r/Stance-Detection-with-Explanations-481E/.

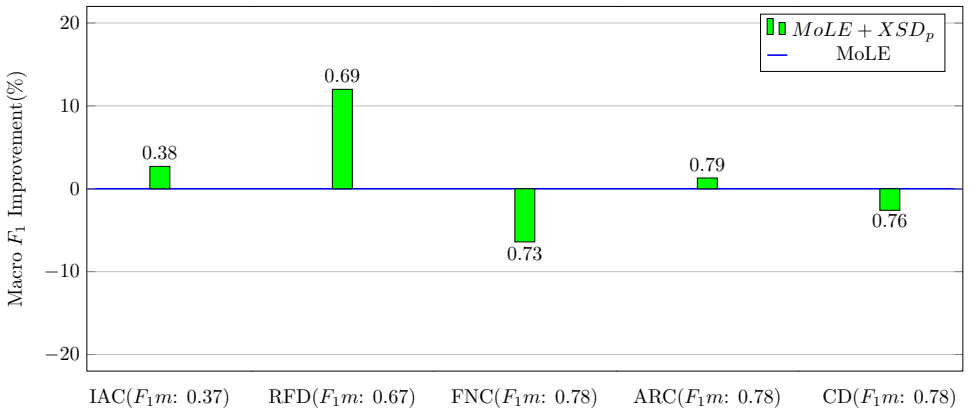



**Figure 8**
Macro $F_1(F_1m)$ score comparison of our approach with pruning ($XSD_p$) and *MoLE* as $Model_{stance}$ w.r.t the best-performing baseline, marked in blue as (*MoLE*) on long (IAC, RFD, FNC) and short (ARC, CD) article datasets. The result shown is for two stance classes, "pro" and "con," numbers in parentheses are the absolute ($F_1m$) scores obtained by *MoLE*. The numbers on top of each bar are the $F_1m$ score obtained by our approach.

**Table 6**
Pairwise Macro $F_1(F_1m)$ score comparison on five datasets while using *MTDNN* and *HSD* as $Model_{stance}$.

| Dataset | *MTDNN* | *MTDNN* + $XSD_p$ | *HSD* | *HSD* + $XSD_p$ |
|---|---|---|---|---|
| IAC | 0.55 | 0.53 | 0.34 | 0.45 |
| RFD | 0.52 | 0.52 | 0.4 | 0.4 |
| FNC | 0.8 | 0.78 | 0.85 | 0.81 |
| ARC | 0.74 | 0.73 | 0.69 | 0.71 |
| CD | 0.77 | 0.76 | 0.83 | 0.83 |

on-par to superior performance on four out of five datasets, achieving a substantial performance gain of 32% on the IAC dataset. These findings reaffirm our claim that our proposed approach for stance explanation can serve as a complementary and effective extension to any stance detection model.

### 6.2 Impact of the Pruning Strategy

In this section, we delve into the advantages of our pruning strategy, aiming to address two crucial questions: (1) Does the implementation of our pruning strategy improve the performance of our architecture compared to no pruning? and (2) How much does pruning reduce the size of stance explanations?

Table 7a presents a comparison of our proposed approach's performance with and without pruning across all five test datasets. For all the experiments regarding the evaluation of the pruning strategy, we used *MoLE* as $Model_{stance}$. The pruning strategy exhibits performance improvements in four of the five datasets, with the most significant gain observed in the FNC dataset (+15.9%). However, for the Createdebate dataset,

**Table 7**
Impact of pruning.

| Dataset | Macro F1 $XSD_{np}$ | Macro F1 $XSD_{p}$ |
|---|---|---|
| IAC | 0.36 | 0.38 |
| RFD | 0.67 | 0.69 |
| FNC | 0.63 | 0.73 |
| ARC | 0.77 | 0.79 |
| CD | 0.76 | 0.76 |

(a) Macro F1 score comparison between our approach, without pruning ($XSD_{np}$) and with pruning ($XSD_{p}$)

| $sim_{thres}$(RFD Dataset) | # Sentences pruned | Macro F1 |
|---|---|---|
| 0.2 | 34.61% | 0.69 |
| 0.3 | 55.19% | 0.68 |
| 0.4 | 77.37% | 0.67 |

(b) Number of sentences pruned and Macro F1 score comparison for different similarity thresholds ($sim_{thres}$) on the RFD dataset

pruning does not lead to a significant performance boost. We hypothesize that this lack of improvement can be attributed to the dataset's predominantly short article lengths, limiting the benefits derived from pruning.

To further investigate the benefits of the pruning strategy, we assess the advantages for readers in terms of the number of sentences pruned. Table 7b demonstrates the effect of varying the similarity threshold ($sim_{thres}$) from 0.2 to 0.4 for the RFD dataset. We report the average number of pruned sentences and the corresponding change in the Macro F1 score. Even with a similarity threshold of 0.2, we can prune nearly 35% of the sentences on average. As we increase the value of $sim_{thres}$, the number of pruned sentences rapidly escalates, with 77% sentences pruned at $sim_{thres} = 0.4$. Notice that the drop in F1 score at high similarity thresholds is modest. This drop can be attributed to the increased risk of losing important arguments from articles that possess a stance but are less similar to the claim. By and large, these results demonstrate the positive impact of our pruning strategy on model performance and highlight the potential time-saving benefits for readers.

### 6.3 Stance Explanation in Relation to First-*k* Article Sentences

In this task, we define stance explanations as a set of sentences that play a pivotal role in determining the author's stance towards a claim. In this section, our objective is to explore the relationship between stance explanations and their position within the article. Specifically, we aim to determine the frequency at which an explanation comprising $k$ sentences can be generated solely from the first $k$ sentences of the article.

To conduct our investigation, we designed an experiment utilizing the RFD dataset and *MoLE* as $Model_{stance}$. We computed the number of cases in which an explanation consisting of $k$ sentences aligns exactly with the initial $k$ sentences of the article. We varied the value of $k$ from 3 to 5. The results of our experiment indicate that, for a 3-sentence explanation, the first 3 sentences of the article are present in approximately 22% of cases. Similarly, for a 4-sentence explanation, the first 4 sentences are present in approximately 15% of cases. Lastly, for a 5-sentence explanation, the first 5 sentences are present in only around 13% of cases. We observed the same trend in other datasets as well. These findings underscore the importance of discourse parsing in uncovering the most crucial arguments in an article.

### 6.4 Experiment with All Classes

In this section, we explain in detail how we utilize our approach for predicting stance for all stance classes, "pro," "con," and "other." We first choose a stance predictor

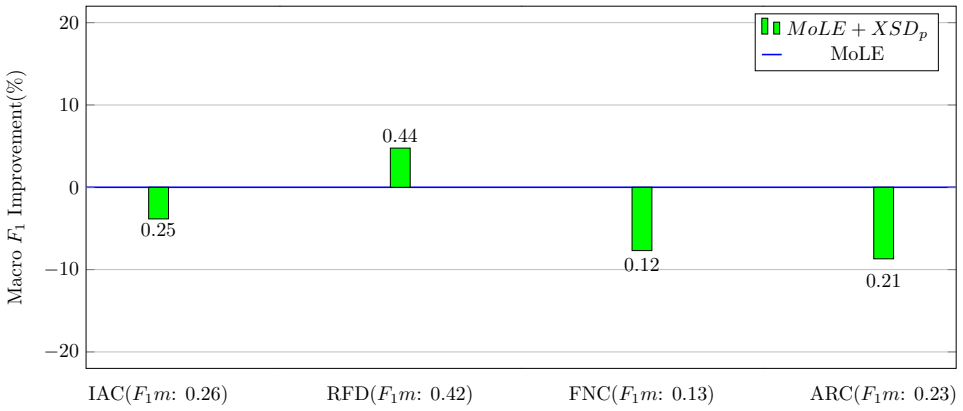



**Figure 9**
Macro $F_1(F_1m)$ score comparison of our approach with pruning ($XSD_p$) and *MoLE* as $Model_{stance}$ w.r.t. the best-performing baseline, marked in blue as (*MoLE*) on long (IAC, RFD, FNC) and short (ARC, CD) article datasets. The result shown is for three stance classes, "pro," "con," and "other"; numbers in parentheses are the absolute ($F_1m$) scores obtained by *MoLE*. The numbers on top of each bar are the $F_1m$ score obtained by our approach.

model (*MoLE*) that was trained on only "pro" and "con" classes and utilize it to make predictions for each claim-sentence (or claim-EDU) pair in a stance tree (as explained in subsection 4.3). Then, we use the DST theory (as explained in subsection 4.6) and compute global prediction and global confidence for each node in the tree. To enable our system to predict the "other" class, we use a simple heuristic rule. In the first stage of our experiment (subsection 6.1), we treated stance detection as a 2-class classification problem; the class with the highest global confidence between the two was the winner (in Figure 2 [bottom row], the global confidences in the "pro" and "con" classes were, respectively, 0.95 and 0.05, hence the final stance was determined to be "pro"). In the second stage, we introduce the following rule: If the global confidence of the winning class is less than a set threshold θ, then we ignore the already aggregated global stance ("pro" or "con") and reclassify it to be of the "other" class.[19]

By adding the capability to predict claim-article pairs that do not belong to "pro" or "con," we ensure that our model is compatible with any stance detection method that is agnostic to specific dataset variations. By incorporating this flexibility, our system maintains its practicality and adaptability to different datasets, ensuring its broader applicability and robustness in stance detection tasks.[20]

Moreover, as demonstrated in Figure 9, our approach exhibits comparable performance on three out of four stance detection datasets (excluding CD) while predicting all stance classes. This finding strongly supports the robustness of our approach, further bolstering the credibility of our claims. By successfully accommodating all stance classes and consistently achieving on-par performance, our approach represents a valuable

19 In this work, we set θ to be 0.52. Since we focus only on the global confidence of the winning class during this experiment and our $Model_{stance}$ is trained on two classes, θ must be greater than 0.5.

20 While predicting on all classes, for fair evaluation, the baseline is also trained on 2 classes and we use the same heuristic rule to predict "other" class.

contribution to stance detection by endowing stance detection systems with argument-based explanations, showcasing its reliability and effectiveness across diverse datasets.

## 7. Conclusions and Future Work

This work explores explanation-based strategies for stance detection in claim-article settings. Leveraging rhetorical parsing, we construct a stance tree that effectively combines evidence using the Dempster-Shafer Theory. Unlike prior works on stance explanations, our approach returns explanations in the form of arguments supporting the stance of an article toward an input claim. Our explanation-based approach yields high-quality explanations, preferred over those generated by state-of-the-art extractive summarizer models while showing superior or on-par performance compared to top-performing models across multiple datasets. A noteworthy aspect of our proposed architecture is its seamless integration capability with any stance detection model, thanks to the incorporation of rhetorical parsing and the unsupervised nature of explanation generation. This feature sets our work apart from existing approaches, making our system versatile and adaptable to any stance detection model. In future investigations, we intend to explore more sophisticated pruning strategies to further enhance the efficiency and effectiveness of our approach. Additionally, we aim to extend the applicability of our system to other NLP tasks, such as claim verification and sentiment analysis.

## 8. Limitations

In this work, we focus on the English language only. Working with other languages would require extensive experiments on the discourse parser and segmenter applicable to those languages. Our experiments show that pruning improves the model performance and the quality of the generated explanations. Currently, our pruning strategy to differentiate arguments from non-stance-taking sentences in the stance tree relies on an external model ($Model_{sim}$), which complicates the pipeline. In our conducted human study, we present turkers with the claim and the explanations generated by three different systems. It is important to acknowledge that we do not provide the articles alongside this information. Consequently, if all three methods happen to miss a crucial argument within the article, the turkers have no means of detecting these omissions. However, it is worth highlighting that our study primarily centers on establishing relative rankings among different systems rather than focusing on absolute numerical assessments. In scenarios where a pivotal argument might go unnoticed by all evaluated systems, the choice of a ranking-based evaluation strategy is less likely to be substantially impacted, as it pertains more to the relative performance of the systems rather than their absolute accuracy in capturing every nuance within the articles.

## 9. Ethical Considerations

**Ethical Consideration Associated with the Task Formulation**: The formulation of our stance explanation approach entails a comprehensive examination of the ethical implications surrounding the task. We acknowledge the importance of privacy and consent in handling sensitive information, especially when exposing an author's stance on controversial topics. By revealing the pivotal arguments from an article, this work aims to benefit various individuals and groups, such as researchers and academics, journalists and fact-checkers, debaters and public speakers, and so forth. Undoubtedly, a system devised for the purpose of determining and elucidating an author's stances,

given controversial claims and authored articles, is susceptible to multiple forms of misuse. Given the recent strides in generative AI, malicious entities have the potential to create articles containing meticulously crafted arguments to manipulate the system's stance determination process, thereby disseminating misinformation or disinformation. In politics, this system could be exploited to strategically influence public sentiment by amplifying specific stances and arguments while marginalizing others, potentially swaying electoral outcomes and public perceptions. Furthermore, in the absence of vigilant moderation, the system might inadvertently produce explanations incorporating hate speech or offensive language from user-contributed articles, thereby facilitating the proliferation of harmful content.

**Ethical Consideration Associated with the Datasets**: The creation and use of the two new stance detection datasets necessitate vigilant ethical considerations in data collection and usage. To ensure responsible practices, we took great care in collecting data from online debate platforms, adhering strictly to the platforms' terms of service and legal requirements. Moreover, in cases where the datasets contained personal information or identifiable data, we prioritized the anonymization of such data to safeguard individual privacy. Pertaining to the actual process of dataset creation, we committed ourselves to comprehensively educating all contributors involved, meticulously furnishing them with a diverse array of enlightening examples encompassing both stance detection and explanation. Beyond these overarching ethical considerations, we have been steadfast in our commitment to conduct this study ethically. This includes transparent communication with turkers to elucidate the purpose and nature of data utilization, as well as ensuring fair compensation for their valuable contributions at a rate reflective of reasonable hourly wages.

**Ethical Consideration Associated with the Proposed Approach**: Our proposed approach for stance explanation entails ethical responsibilities regarding transparency, accountability, and potential misuse. We emphasized the need for clear explanations in our approach, ensuring that readers and stakeholders comprehend the process leading to stance predictions. The dominant source of potential bias within our approach can likely be traced to the caliber of the stance detection model ($Model_{stance}$). Although prior studies have illustrated the presence of bias in stance detection models (Yuan, Zhao, and Qin 2022; Kaushal, Saha, and Ganguly 2021; Schiller, Daxenberger, and Gurevych 2021), the adaptable architecture of our proposed framework inherently empowers conscientious users to select the most contemporary and impartial stance detection model within our methodology.

## Appendix A. Similarity Pruning Algorithm

**Algorithm 1** Pruning DDT based on Similarity

```
1:  procedure SIM-PRUNE(DDT, SL)
2:      Repeat from the second last layer to top layer of the DDT
3:      Repeat for every node in a layer
4:      parent ← node
5:      children ← Children(parent);
6:      for all child in children do
7:          if child ∉ SL then
8:              children.remove(child);
9:              DDT.remove(child);
10:     if isEmpty(children) and parent ∉ SL then
11:         DDT.remove(parent);
12:     else
13:         if parent ∉ SL then
14:             if not (∃isSatellite(children) and isNucleus(parent)) then
15:                 lmn ← LMN(children)
16:                 if lmn then
17:                     DDT.replace(parent, lmn)
18:                 else
19:                     lms ← LMS(children)
20:                     DDT.replace(parent, lms)
```

Algorithm 1 describes the steps for similarity-based pruning. We use $Model_{sim}$ to compute the cosine similarity between claim and EDU (or sentence) and add only those to *SL* whose similarities were higher than a similarity threshold. Algorithm 1 prunes less-similar EDUs (or sentences) as we progress to the root of the DDT from the leaves. For every parent, if it is present in *SL*, then we simply remove its children which are not present in *SL* (line 6-9). However, if the parent is not present in *SL*, then we first prune children that are not in *SL*. If this leads to a state when the parent has no children left, we also remove the parent from DDT (line 10-11). Finally, we handle two special cases:

- If, after pruning, one of the surviving children is a satellite and the parent is a nucleus, then we do not prune the parent. We follow this strategy because the nucleus is more informative than satellites.
- In any other case, we replace the parent with its left most nucleus (*lmn*) child or with the left most satellite (*lms*) child if *lmn* does not exist (line 15-20).

## Appendix B. Survey Guidelines

Before presenting the explanations to the turkers and initiating the evaluation process, we ensure that they receive detailed instructions. These instructions serve as a guide to familiarize the turkers with the task of stance detection and explanation. We provide three illustrative examples to introduce the task. Following this, we describe the evaluation criteria in depth. Additionally, we offer a comprehensive example consisting of various snippets. Each snippet represents explanations generated from either Matchsum, T5, and $XSD_{np}$ (in the first survey) or $XSD_{p}$ (in the second survey). In the example, we provide arguments to explain why a particular snippet receives a high coverage score but a low non-redundancy score. We emphasize the importance of understanding these scores for the evaluation process. By providing these detailed instructions, we ensure that the turkers have a clear understanding of the task, the evaluation criteria, and the presentation format of the explanations.

## Appendix C. Example Question for First Survey

**Table C.1**
Example of a question shown in the first survey where we compare explanations generated by *T*5, $XSD_{np}$, and *MS*.

**Claim:** Foreign Language Classes should Be Mandatory in College.
Stance Label: **con**

**Snippet 1**
1. high school students look forward to college as a place where they are free to choose what they want to study.
2. learning thinking skills, or even communication, isn't unique to foreign language classes, says physicist edward mccartney, who studied in laos for 11 years, can make real fluency seem unattainable.
3. the level that students reach even after years in an overseas language course is even clearer.
4. if you are studying abroad, you can learn more through your other classes or life.com ... and?
5. school 'the way they can't, sydney d' can tyranny and other i's - comma neo yayote obama, and learning new language feming and the way he thinks that learning the language can help resads can get the u.s. foreign students can understand the world & p ccs or other courses based on thai language and learn w r and your own!
6. learning l g.

**Snippet 2**
1. High school students look forward to college as a place where they are free to choose what they want to study and what career path they want to take. (con)
1.1 While learning a foreign language theoretically presents a valuable opportunity to communicate or think in a different way, a foreign language requirement is problematic. (con)
1.1.1. First, it assumes that the same process of learning new "languages" and ways of thinking can't be accessed by simply studying a different discipline; and second, that students will learn a language well enough to actually experience a culture. (con)
...
1.2. Learning thinking skills, or even communication, isn't unique to foreign language classes. (con)
1.2.1. While it may be true that argumentation differs across cultures (in Laos, arguments may be less direct than in India), it is also true that a physicist argues a point differently from a literary theorist. (con)
...

**Snippet 3**
1. High school students look forward to college as a place where they are free to choose what they want to study and what career path they want to take.
2. While learning a foreign language theoretically presents a valuable opportunity to communicate or think in a different way, a foreign language requirement is problematic.
3. First, it assumes that the same process of learning new "languages" and ways of thinking can't be accessed by simply studying a different discipline; and second, that students will learn a language well enough to actually experience a culture.
...
...
12. Taking courses other than foreign languages, though maybe not as obviously cultural, gives students a chance to explore new or foreign ideas with the basics of communication already in place.
13. I want to be challenged in the way I think about an issue, not in the way I express that thinking.

In Table C.1, we show an example that was shown to the turkers during the first survey. The example contains a claim, the stance label, and stance explanations generated from T5, $XSD_{np}$, and MS systems, respectively.

## Appendix D. Survey Result for Small and Medium Bins

The survey results for the small and medium bins are presented in Figure D.1 and Figure D.2, respectively. In these figures, we compare the average scores obtained by our approach with pruning ($XSD_p$) alongside Matchsum and T5 over four criteria, including informativeness, coverage, non-redundancy, and overall quality. In the small bin, $XSD_p$ demonstrates significant improvements compared to Matchsum, while T5 performs the worst. Specifically, $XSD_p$ exhibits a 13.2% enhancement in terms of informativeness, a 2.3% improvement in coverage, a 14.6% improvement in non-redundancy, and a 12.3% improvement in overall quality. Moving to the medium bin, $XSD_p$ continues to outperform Matchsum across all evaluation criteria, while T5 continues to perform the worst. The results indicate a substantial 15.1% improvement in informativeness, a 3.71% improvement in coverage, a notable 17.9% improvement in non-redundancy, and a 14.6% improvement in overall quality. These findings provide strong evidence that, as we progress from the small bin to the medium bin, turkers consistently exhibit a preference for the explanations generated by $XSD_p$ over Matchsum.

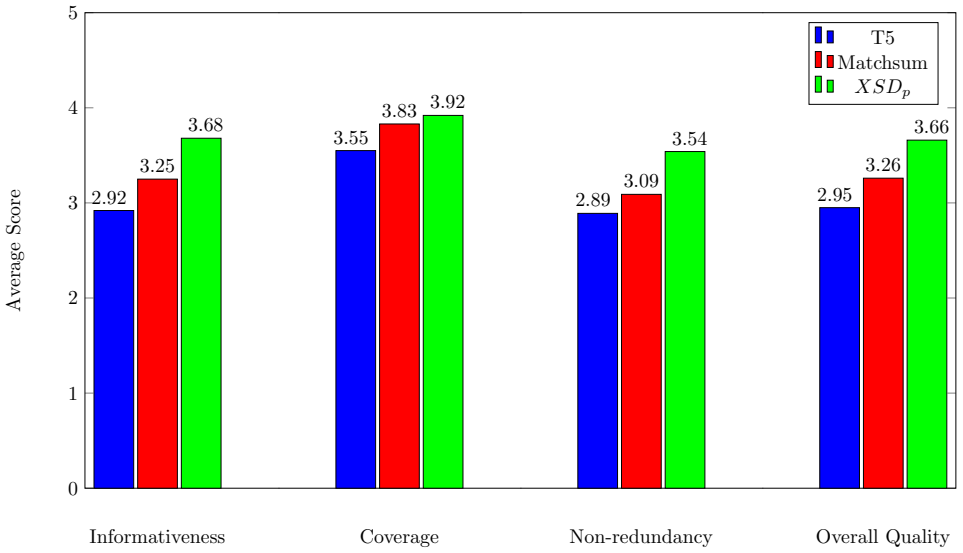



**Figure D.1**
Comparison of the obtained average score (range is 0 to 5) for the stance explanations generated from our system with pruning ($XSD_p$), *T5*, and Matchsum (*MS*) across the four evaluation criteria in the ***small*** bin.

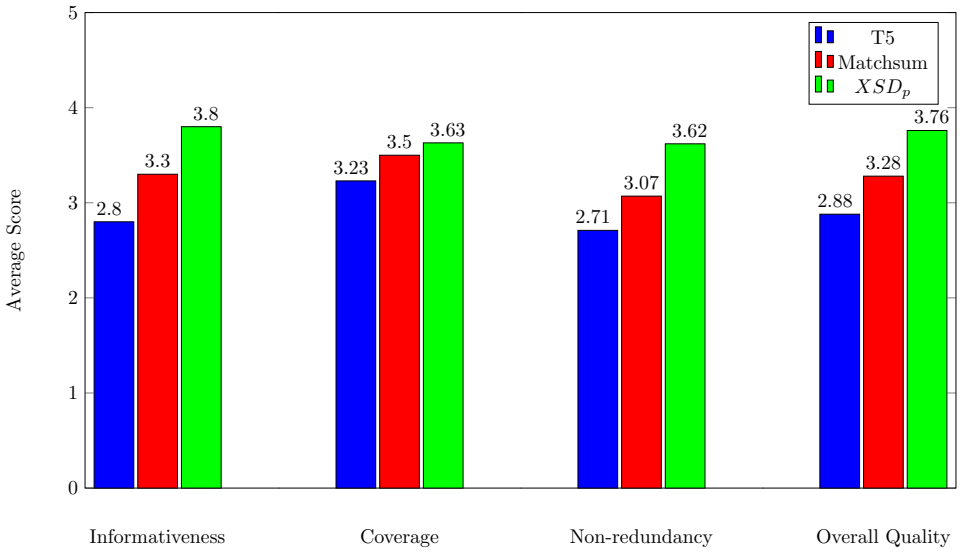



**Figure D.2**
Comparison of the obtained average score (range is 0 to 5) for the stance explanations generated from our system with pruning ($XSD_p$), *T5*, and Matchsum (*MS*) across the four evaluation criteria in the ***medium*** bin.

## Appendix E. Baseline Implementations

In this study, we used existing codes for all the baselines, adhering to the parameter settings proposed in the original papers. Notably, we ensured that the *HSD* baseline was trained and tested on individual datasets. Conversely, for *MoLE* and *MTDNN*, we integrated CD and RFD with the original pool of datasets used during training.

Regarding the *MoLE* baseline, which was initially trained on 16 stance detection datasets, we faced challenges in obtaining two datasets, namely, Rumor (Qazvinian et al. 2011) and MTSD (Sobhani, Inkpen, and Zhu 2017). Despite our efforts to acquire the data, the authors remained unresponsive or unwilling to share the datasets. Moreover, Twitter's recent policy change to replace the free API with paid tiers further hindered our access to the required data (`https://www.theverge.com/2023/2/2/23582615/twitter-removing-free-api-developer-apps-price-announcement`).

For the *HSD* baseline, we implemented the codes necessary for computing textual features, such as cosine similarity, overlap coefficient, tf-idf score, Word2vec similarity, and more. This step was essential as these features were missing from the authors' original implementation.

## Appendix F. Comparison of Aggregation Policies

We compare the DST policy against two other aggregation policies (majority and average) while fixing the model to be MT-DNN$_{\text{MDL}}$. In the majority policy, during evidence aggregation, if the number of pro children is greater than the number of con children for any subtree, we mark the parent as having a pro stance and vice versa. For the average

policy, we take a similar action when the mean of the confidence of the pro children exceeds that of the con children. In both cases, the global confidence of the parent is computed as the mean of the confidence of the children in the winning stance. Our experiments report a higher Macro $F_1$-score for DST (**0.76**) compared with the majority (0.73) and average (0.74) policy.

## Appendix G. Result Analysis on EDU Level

Our hypothesis is that sentences play a better role in capturing arguments than EDUs because of the latter's much shorter length. To verify this, we compare the prediction from EDU-based *Stance Tree* against the baselines on the CD dataset. With respect to the best performing Schiller's architecture ($MT - DNN$), we observe a 10% deterioration in Macro $F_1$ score for our approach with pruning. Also, adding the (dis)agreement detection module does not have an impact on the overall performance. Based on these results, we do not conduct further experiments using EDU-based *Stance Tree* on other datasets.

## Appendix H. Training Baseline on All Classes

In subsection 6.4, we presented results where the baseline model (*MoLE*) is trained on two classes ("pro" and "con") and is used to predict stances for three classes ("pro," "con," and "other") via a heuristic rule. In this section, we present the results when the baseline model is trained on all three classes. Specifically, we have conducted experiments comparing the following two settings:

1. $MoLE_{3-3}$: In this setting, the baseline model (e.g., *MoLE*) is trained and tested on all three classes ("pro," "con," and "other").
2. $XSD_{2-3}$: the $Model_{stance}$ in our approach (i.e., $XSD_p$) is trained on 2 classes ("pro"and "con") and then tested on all 3 classes.

While comparing between $MoLE_{3-3}$ and $XSD_{2-3}$, we made the following observations:

1. For ARC and FNC datasets, $XSD_{2-3}$ attains a superior accuracy score for the "pro" and "con" classes compared to $MoLE_{3-3}$. This can be attributed to the fact that both of these datasets are dominated by the number of examples from the "other" class. Therefore, in the $MoLE_{3-3}$ setting, by training on all classes, a large number of "pro" and "con" examples tend to get misclassified when we introduce data from the "other" class.
2. $MoLE_{3-3}$ setting achieves a superior accuracy for the "other" class compared to the $XSD_{2-3}$ approach for three out of four datasets. For example, In the IAC dataset, accuracy for the "other" class improves from 3.96 to 95.62 when we compare between $XSD_{2-3}$ and $MoLE_{3-3}$. We observe similar trends in FNC and IAC datasets too.

It is also important to emphasize that the content of the "other" class is not guaranteed to "lie between" the contents of the "pro" and "con" classes as it includes everything that is not pro and not con. For example, an article can take an "other" stance with respect to a claim in all the following scenarios: (a) the article contains both pro segments and con segments towards a claim and reaches a neutral conclusion; (b) the article merely discusses the claim without taking any stance; (c) the article is completely unrelated to the claim. Our goal in this work is to provide explanations for "pro" and "con." Providing explanations for "other" is far more complex and does not easily lend itself to DST's reasoning and propagation.

## Appendix I. Comparison of Computational Time

In our approach (*XSD*), as explained in subsection 4.3, we rely on a stance detection model capable of predicting the stance for a claim-sentence pair. However, unlike existing models that make a single prediction for a given input pair, our approach goes a step further by segmenting the article into Elementary Discourse Units (EDUs) or sentences and computes the stance for each claim-EDU (or claim-sentence) pair. This naturally raises the question: How does our approach compare to a baseline model in terms of computation time?

While hypothetically *XSD* can consume more energy/wall time compared to typical baselines, we do not consider it to be significant for the following reasons:

1. Firstly, it is important to note that our use of the stance predictor model ($Model_{stance}$) within the *XSD* framework involves a training process that needs to be executed only once. The training data for $Model_{stance}$ is consistent across all baseline models, thereby eliminating any variance in energy consumption or wall clock time during the training phase between *XSD* and the baseline models. Notably, the distinction arises solely during inference, where *XSD*'s predictive model accommodates claim-EDU or claim-sentence pairs, while the baseline models handle single claim-article pairs. Although this leads to a larger number of predictions for *XSD* during testing, we emphasize that the trained model is reused, mitigating the impact on inference time.
2. Secondly, combining the predictions from *XSD* to establish a global decision involves the iterative application of the Dempster-Shafer rule on the discourse dependency tree. This rule is characterized by straightforward mathematical operations, making it remarkably fast with negligible influence on energy consumption or wall clock time.

To provide empirical evidence of this, we conducted experiments on the ARC and FNC datasets using *MoLE* as the baseline and compared it with our approach ($XSD_p$). For the ARC dataset, *MoLE* and $XSD_p$ take approximately 4.45 minutes and 5.26 minutes, respectively. For the FNC dataset, *MoLE* and $XSD_p$ take approximately 6.55 and 11.27 minutes, respectively. In both scenarios, most of the increase in wall clock time for $XSD_p$ can be attributed to a large number of predictions during inference time as illustrated in the first point above. Furthermore, it's important to underscore that this small increase in computational time is justified by the fact that our approach pro-

vides stance explanations, at a fine granularity unlike existing models. This additional capability enhances the interpretability and transparency of our model, which can be invaluable in many real-world applications.

## References


Agerri, Rodrigo, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Allaway, Emily and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*. https://doi.org/10.18653/v1/2020.emnlp-main.717

Allaway, Emily, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 4756–4767. https://doi.org/10.18653/v1/2021.naacl-main.379

Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 1–9.

Augenstein, Isabelle, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*. https://doi.org/10.18653/v1/D16-1084

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *arXiv preprint cmp-lg/9602004*.

Cignarella, Alessandra Teresa, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. Sardistance@ evalita2020: Overview of the task on stance detection in Italian tweets. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–10. https://doi.org/10.4000/books.aaccademia.7084

Clark, Thomas, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi, and Nigel Collier. 2021. Integrating transformers and knowledge graphs for Twitter stance detection. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 304–312. https://doi.org/10.18653/v1/2021.wnut-1.34

Conforti, Costanza, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101. https://doi.org/10.18653/v1/2020.findings-emnlp.365

Conforti, Costanza, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2022. Incorporating stock market signals for Twitter stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4074–4091. https://doi.org/10.18653/v1/2022.acl-long.281

Del Tredici, Marco, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. *arXiv preprint arXiv:1909.00412*. https://doi.org/10.18653/v1/D19-1477

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dey, Kuntal, Ritvik Shrivastava, and Saroj Kaushik. 2017. Twitter stance detection—A subjectivity and sentiment polarity inspired two-phase approach. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 365–372. https://doi.org/10.1109/ICDMW.2017.53

Du, Jiachen, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3988–3994. https://doi.org/10.24963/ijcai.2017/557

Dulhanty, Chris, Jason L Deglint, Ibrahim Ben Daya, and Alexander Wong. 2019. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *arXiv preprint arXiv:1911.11951*.

Fang, Wei, Moin Nadeem, Mitra Mohtarami, and James Glass. 2019. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 13–19. https://doi.org/10.18653/v1/D19-6603

Faulkner, Adam. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *The Twenty-Seventh International Flairs Conference*, pages 174–179.

Ghosh, Subrata, Konjengbam Anand, Sailaja Rajanala, A. Bharath Reddy, and Manish Singh. 2018. Unsupervised stance classification in online debates. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 30–36. https://doi.org/10.1145/3152494.3152497

Glandt, Kyle, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)*, volume 1. https://doi.org/10.18653/v1/2021.acl-long.127

Graham, Yvette. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137. https://doi.org/10.18653/v1/D15-1013

Grootendorst, Maarten. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Gunhal, Pranav, Aditya Bashyam, Kelly Zhang, Alexandra Koster, Julianne Huang, Neha Haresh, Rudransh Singh, and Michael Lutz. 2022. Stance detection of political tweets with transformer architectures. *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. pages 658–663. https://doi.org/10.1109/ICTC55196.2022.9952951

Hanselowski, Andreas, P. V. S. Avinesh, Benjamin Schiller, and Felix Caspelherr. 2017. Description of the system developed by team athene in the fnc-1. *Fake News Challenge*.

Hanselowski, Andreas, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*.

Hardalov, Momchil, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028. https://doi.org/10.18653/v1/2021.emnlp-main.710

Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762. https://doi.org/10.3115/v1/D14-1083

Hayashi, Katsuhiko, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for rst discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136. https://doi.org/10.18653/v1/W16-3616

He, Zihao, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. *arXiv preprint arXiv:2204.03839*. https://doi.org/10.18653/v1/2022.wassa-1.7

Hewett, Freya, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103. https://doi.org/10.18653/v1/W19-4512

Hirao, Tsutomu, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013*

*Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520.
Hou, Shengluan and Ruqian Lu. 2020. Knowledge-guided unsupervised rhetorical parsing for text summarization. *Information Systems*, 94:101615. https://doi.org/10.1016/j.is.2020.101615
Jain, Kushal, Fenil R. Doshi, and Lakshmi Kurup. 2020. Stance detection using transformer architectures and temporal convolutional networks. https://doi.org/10.1007/978-981-15-4409-5_40
Jayaram, Sahil and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554. https://doi.org/10.18653/v1/2021.emnlp-main.450
Jia, Ruipeng, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631. https://doi.org/10.18653/v1/2020.emnlp-main.295
Kaushal, Ayush, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889. https://doi.org/10.18653/v1/2021.naacl-main.303
Kobbe, Jonathan, Ioana Hulpuş, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60. https://doi.org/10.18653/v1/2020.emnlp-main.4
Kraus, Mathias and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79. https://doi.org/10.1016/j.eswa.2018.10.002
Krejzl, Peter, Barbora Hourová, and Josef Steinberger. 2017. Stance detection in online discussions. *arXiv preprint arXiv:1701.00504*.
Lefevre, Eric, Olivier Colot, and Patrick Vannoorenberghe. 2002. Belief function combination and conflict management. *Information Fusion*, 3(2):149–162. https://doi.org/10.1016/S1566-2535(02)00053-2
Li, Sujian, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35. https://doi.org/10.3115/v1/P14-1003
Li, Yingjie and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305. https://doi.org/10.18653/v1/D19-1657
Liang, Bin, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91. https://doi.org/10.18653/v1/2022.acl-long.7
Liu, Rui, Zheng Lin, Huishan Ji, Jiangnan Li, Peng Fu, and Weiping Wang. 2022. Target really matters: Target-aware contrastive learning and consistency regularization for few-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6944–6954.
Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*. https://doi.org/10.18653/v1/P19-1441
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
Liu, Zhengyuan, Ke Shi, and Nancy F. Chen. 2020. Multilingual neural RST discourse parsing. *arXiv preprint arXiv:2012.01704*. https://doi.org/10.18653/v1/2020.coling-main.591
Loria, Steven. 2018. TextBlob documentation. *Release 0.15*, 2.
Luo, Ling, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

*Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3033–3043. https://doi.org/10.18653/v1/D19-1300

Luo, Yun, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Mao, Yuning, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4941–4957. https://doi.org/10.18653/v1/2020.acl-main.445

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. https://doi.org/10.18653/v1/S16-1003

Mohtarami, Mitra, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*. https://doi.org/10.18653/v1/N18-1070

Murphy, Catherine K. 2000. Combining belief functions when evidence conflicts. *Decision Support Systems*, 29(1):1–9. https://doi.org/10.1016/S0167-9236(99)00084-6

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759. https://doi.org/10.18653/v1/N18-1158

Pick, Ron Korenblum, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. STEM: Unsupervised STructural EMbedding for Stance Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11174–11182. https://doi.org/10.1609/aaai.v36i10.21367

Popat, Kashyap, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. STANCY: Stance classification based on consistency cues. *arXiv preprint arXiv:1910.06048*. https://doi.org/10.18653/v1/D19-1675

Pu, Dongqi, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. *arXiv preprint arXiv:2305.16784*. https://doi.org/10.18653/v1/2023.acl-long.306

Qazvinian, Vahed, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Association for Computational Linguistics, Edinburgh, Scotland, UK.

Qian, Shenbin, Constantin Orăsan, Diptesh Kanojia, Hadeel Saadany, and Félix Do Carmo. 2022. SURREY-CTS-NLP at WASSA2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 271–275. https://doi.org/10.18653/v1/2022.wassa-1.29

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rakholia, Neel and Shruti Bhargava. "Is it true?"—Deep learning for stance detection in news. *Fake News Challenge*.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, https://doi.org/10.18653/v1/2020.emnlp-main.365

Respall, Victor Massague and Leon Derczynski. 2017. Stance detection in Catalan and Spanish tweets. *recall*, 1(2):1.

Riedel, Benjamin, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat

baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Ruder, Sebastian, John Glover, Afshin Mehrabani, and Parsa Ghaffari. 2018. 360 stance detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 31–35. https://doi.org/10.18653/v1/N18-5007

Samih, Younes and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646. https://doi.org/10.18653/v1/2021.eacl-main.227

Schiller, Benjamin, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*. 1–13. https://doi.org/10.1007/s13218-021-00714-w

Sean Baird, Doug Sibley and Yuxi Pan. 2017. Talos System Description. https://blog.talosintelligence.com/talos-fake-news-challenge/. Accessed: 2010-09-30.

Sentz, Kari, Scott Ferson, et al. 2002. *Combination of Evidence in Dempster-Shafer Theory*, volume 4015, Sandia National Laboratories Albuquerque. https://doi.org/10.2172/800792

Sepúlveda-Torres, Robiert, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. 2021. Exploring summarization to enhance headline stance detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 243–254. https://doi.org/10.1007/978-3-030-80599-9_22

Shafer, Glenn. 1976. *A Mathematical Theory of Evidence*. Princeton University Press. https://doi.org/10.1515/9780691214696

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

Sobhani, Parinaz, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Somasundaran, Swapna and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. https://doi.org/10.3115/1687878.1687912

Somasundaran, Swapna and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Spadaro, Daniel C., Lawrence A. Robinson, and L. Tracy Smith. 1980. Assessing readability of patient information materials. *American Journal of Hospital Pharmacy*, 37(2):215–221. https://doi.org/10.1093/ajhp/37.2.215

Stab, Christian and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659. https://doi.org/10.1162/COLI_a_00295

Sun, Qingying, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Voorbraak, Frans. 1991. On the justification of Dempster's rule of combination. *Artificial Intelligence*, 48(2):171–197. https://doi.org/10.1016/0004-3702(91)90060-W

Walker, Marilyn A., Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of LREC*, pages 812–817.

Wang, Yizhong, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.

https://doi.org/10.18653/v1/P17-2029

Wang, Yizhong, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. *arXiv preprint arXiv:1808.09147*. https://doi.org/10.18653/v1/D18-1116

Wei, Penghui, Wenji Mao, and Daniel Zeng. 2018. A target-guided neural memory model for stance detection in Twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. https://doi.org/10.1109/IJCNN.2018.8489665

Wei, Wan, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388. https://doi.org/10.18653/v1/S16-1062

Wu, Minghao, Lizhen Qu, George Foster, and Gholamreza Haffari. Improving document-level neural machine translation with discourse features. *Available at SSRN 4330827*.

Wu, Yuxiang and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 5602–5609. https://doi.org/10.1609/aaai.v32i1.11987

Xu, Ruifeng, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC shared task 4: Stance detection in Chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, Springer, pages 907–916. https://doi.org/10.1007/978-3-319-50496-4_85

Yeung, Andy W. K., Tazuko K. Goto, and W. Keung Leung. 2018. Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae. *Frontiers in Human Neuroscience*, 12:308. https://doi.org/10.3389/fnhum.2018.00308, PubMed: 30158861

Yuan, Jianhua, Yanyan Zhao, Yanyue Lu, and Bing Qin. 2022. SSR: Utilizing simplified stance reasoning process for robust stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6846–6858.

Yuan, Jianhua, Yanyan Zhao, and Bing Qin. 2022. Debiasing stance detection models with counterfactual reasoning and adversarial bias learning. *arXiv preprint arXiv:2212.10392*.

Zhang, Qiang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. 2019. From stances' imbalance to their hierarchical representation and detection. In *The World Wide Web Conference*, pages 2323–2332. https://doi.org/10.1145/3308558.3313724

Zhong, Ming, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*. https://doi.org/10.18653/v1/2020.acl-main.552

Zhou, Yiwei, Alexandra I. Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, pages 18–32. https://doi.org/10.1007/978-3-319-68783-4_2