Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1008–1018
Marseille, 11–16 May 2020

Machine-Aided Annotation for Fine-Grained Proposition Types in Argumentation

Yohan Jo[1], **Elijah Mayfield**[1], **Chris Reed**[2], **Eduard Hovy**[1]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Centre for Argument Technology, University of Dundee, UK
yohanj@cs.cmu.edu, elijah@cmu.edu, c.a.reed@dundee.ac.uk, hovy@cmu.edu

Abstract

We introduce a corpus of the 2016 U.S. presidential debates and commentary, containing 4,648 argumentative propositions annotated with fine-grained proposition types. Modern machine learning pipelines for analyzing argument have difficulty distinguishing between types of propositions based on their factuality, rhetorical positioning, and speaker commitment. Inability to properly account for these facets leaves such systems inaccurate in understanding of fine-grained proposition types. In this paper, we demonstrate an approach to annotating for four complex proposition types, namely *normative claims*, *desires*, *future possibility*, and *reported speech*. We develop a hybrid machine learning and human workflow for annotation that allows for efficient and reliable annotation of complex linguistic phenomena, and demonstrate with preliminary analysis of rhetorical strategies and structure in presidential debates. This new dataset and method can support technical researchers seeking more nuanced representations of argument, as well as argumentation theorists developing new quantitative analyses.

Keywords: argumentation theory, proposition types, imbalanced annotation tasks, hybrid annotation systems

1. Introduction

Argument mining is a broad field of computational linguistics that seeks to identify the structure of written and spoken argument and extract meaningful content based on that understanding. But as the domains that we can tackle with NLP grow more diverse, and expand from newswire text to social media and real-world dialogue, we are reaching an inflection point. These domains are not characterized solely by objective statements with clean reporting of facts and details; opinion, hedging, and reported speech are commonplace. In recent years, researchers have found that argument mining pipelines struggle to identify factual content and disambiguate it from fiction, lies, or mere hypotheticals in real-world data (Feng et al., 2012; Thorne et al., 2018). In today's politically charged atmosphere, this poses a challenge for developers of systems like fake news detectors and recommender systems: when algorithmic systems cannot even reliably detect the presence or assertion of facts in statements, how can they address the ethical challenges of deployed machine learning systems at scale (Leidner and Plachouras, 2017; Gonen and Goldberg, 2019)?

This paper introduces new resources for understanding propositions that appear in speech and text, based on the 2016 U.S. presidential debates. We define a fine-grained, four-dimensional annotation schema for how propositions are introduced rhetorically in debates: namely, **normative** statements, statements of **desire**, statements about **future possibility**, and **reported speech**. These proposition types are tied closely to practical reasoning, causal reasoning, and authority claims in argumentation schemes (Walton et al., 2008) and represent varying levels of speaker commitment to individual statements (Lasersohn, 2009).

While these definitions are tractable for reliable human annotators, we find that occurrences in running text are rare and annotation is both difficult and inefficient. In response, we develop a machine learning model with high recall for finding likely candidates for positive labels, and describe a hybrid annotation workflow that boosts the efficiency of human annotators by 39-85% while further improving reliability. Using this process we produce a corpus of annotated propositions to be released alongside this paper. We conclude with a preliminary analysis of how these proposition types are used in political debate and commentary. Our contributions are as follows:

- **A multi-dimensional annotation schema for fine-grained proposition types that are tied to argumentation schemes and speaker commitment.** In our work this schema has been proven to be tractable and robust for both human and automated annotation.
- **A public sample annotated corpus of propositions using that schema, along with full annotation manuals and baseline classification code.**[1] This dataset contains annotated instances of novel proposition types, such as reported speech, and is more than three times larger than comparable recent corpora. All these materials may enable further progress in the community.
- **An effective, efficient, and novel methodology for hybrid machine-aided annotation.** To address logistic challenges with annotating sparse labels in our task, we introduce additional best practices for hybrid human-machine systems for building datasets. This method produces efficient machine filtering, especially of likely negative instances, which covers a large percentage of our corpus. Human annotator time is prioritized on potential positive instances, which are harder to recognize automatically with high precision.

2. Background

2.1. Argument Mining and Proposition Types

Argument mining is an expansive field with many applications. Datasets include the Internet Argument Corpus for

[1]https://github.com/yohanjo/lrec20

online debate on political topics (Walker et al., 2012; Swanson et al., 2015), student argument in course essays (Stab and Gurevych, 2017), and parliamentary debate (Duthie et al., 2016). State-of-the-art results have been produced using a range of methods including random forests (Aker et al., 2017), integer linear programming for constraint-based inference (Persing and Ng, 2016), graph-based methods that focus on relations between claims (Niculae et al., 2017; Nguyen and Litman, 2018), and more recently, end-to-end neural methods (Cocarascu and Toni, 2018; Frau et al., 2019). But these systems struggle to distinguish between distinctions in argumentative strategy that look intuitively obvious to casual observers, instead relying on coarse notions of claims and supports.

Today, automated systems fail to understand the nuanced factuality of these sentences when they appear in argumentation. Perceived factuality of propositions is heavily tied to a speaker's intent (Wentzel et al., 2010); this concept of speakers making claims with only partial certainty or factuality have been collectively studied under the umbrella term of "commitment" to a truth value for claims (Lasersohn, 2009). Naderi and Hirst (2015) give examples of propositions that are not straightforwardly factual, but instead contain propositions deeply embedded in hypotheticals and shifts in tense, beyond the current bounds of today's NLP:

> "*Who among us would dare consider returning to a debate on the rights of women in our society or the rights of visible minorities?*"

> "*How can we criticize China for imprisoning those who practise their religion when we cannot offer protection of religious beliefs in Canada?*"

Later, Haddadan et al. (2018) describe the context-dependent annotation task of identifying premises and claims in political discourse, providing the following sentence from the 1960 Nixon-Kennedy presidential debate:

> "*Communism is the enemy of all religions; and we who do believe in God must join together. We must not be divided on this issue.*"

It turns out ideas are not only factual or fictitious, but lie on a many-dimensional gradient. They can be positioned carefully when making arguments, negotiating, or manipulating a discourse (Potter, 1996), and authors take care to distinguish between claims they know to be true, desires they have for the future, amid other epistemological states of reported knowledge (Walton et al., 2008).

In argumentation theory and communication sciences, propositions are typically divided into three types: fact, value, and policy (Hollihan and Baaske, 2015; Wagemans, 2016). Propositions of *fact* have contents whose truth value is verifiable with empirical evidence, whereas propositions of *value* are subjective judgments. Propositions of *policy* propose that an action be carried out. These types have been extended by prior studies. For instance, Park and Cardie (2018) extended *fact* into *non-experiential fact* and *testimony*, and added *reference*—a text of information source (but not reported speech in itself). Egawa et al. (2019) further added *rhetorical statement*, judgments of *value* using figurative language and discourse structure.

While most prior work extended proposition types based on the needs of the task at hand, our taxonomy has been motivated mainly by argumentation theory. In particular, the argumentation schemes of Walton et al. (2008) are a set of reasoning types commonly used in daily life. Each scheme defines the form of a conclusion and the form(s) of one or more premises. As an example, the scheme of *argument from consequences* is as follows:

> Premise: If *A* is brought about, good consequences will plausibly occur.
> Conclusion: *A* should be brought about.

These schemes have been adopted by many studies as a framework for analyzing reasoning patterns (Song et al., 2017; Nussbaum, 2011). Researchers in computational linguistics have tried to code the schemes, but this task turned out to be very challenging; as a result, annotations have low agreement between annotators (Lindahl et al., 2019) or are available only from experts (Lawrence et al., 2019). But different schemes are associated with different proposition types, and therefore, we speculate that reliably annotating proposition types may ease the annotation of argumentation schemes. The proposition types in this paper are closely related to common argumentation schemes, including practical reasoning, argument from consequence, argument from cause to effect, and argument from expert opinion.

2.2. Efficient Linguistic Annotation

In their overview of argument mining today, Lippi and Torroni (2016) identify three key challenges that limit the field:

1. The subtlety of the task requires more time-consuming and expensive training to achieve high inter-rater reliability, compared to tasks like object detection in computer vision, limiting the size and breadth of corpora available to researchers.
2. Because of the lack of existing data, there are few automation tools available to expedite the annotation of future datasets, leaving the field with too much unsupervised data and not enough labels.
3. The structured nature of claims and premises limits the utility of widely-used classification algorithms.

More recent reviews of the field have made similar observations (Lawrence and Reed, 2019; Janier and Saint-Dizier, 2019). Researchers have suspected that part of the challenge in these problems is data collection and reliable annotation. Collecting span- and sentence-level annotations is a frequently used tool for machine learning researchers seeking to improve their systems. Accurate annotation is time-consuming and expensive, though, and even when funding is available, annotation tasks often require subject matter expertise that comes from either lived experience or extensive training. This problem is exacerbated by rare phenomena, which results in imbalanced datasets in many domains, like emotional crisis or suicidal ideation detection online and in medical records (Pestian et al., 2012; Imran et al., 2016; Losada and Crestani, 2016), rare occurrence of high- and low-end scores in student data in education domains (Woods et al., 2017; Lugini and Litman, 2018), and rare social behaviors in healthcare settings (Mayfield et al., 2013; Carrell et al., 2016). Our annotation also handles rare phenomena,

and using a conventional annotation methodology allows only moderate inter-annotator agreement even after intensive annotator training, reflecting the difficulty of our task.

Many previous papers on text annotation have relied on crowdsourcing, relying on inexperienced editors on services such as Crowdflower and Amazon Mechanical Turk (Snow et al., 2008; Swanson et al., 2015). While this approach works for many common-sense tasks, prior work has shown that achieving high inter-rater reliability with these services is arduous and relies on many strict methodological choices and narrowing of task type (Alonso et al., 2015; Hoffman et al., 2017). When converting real-world phenomena into categorical judgments that can achieve high reliability, nuance is often lost in the name of inter-annotator agreement. This requires researchers to make a trade-off between, on one hand, the expressiveness and fidelity of the linguistic construct they are attempting to capture, and on the other the potential for operationalization and quantification in coding manuals and fully automated systems. Particularly in imbalanced tasks, these choices can have the effect of producing an inaccurate picture of the minority class and producing datasets that are no longer a valid representation of the original construct (Corbett-Davies and Goel, 2018).

To expedite annotation without sacrificing validity, researchers have developed annotation tools that incorporate machine learning (Pianta et al., 2008; Yimam et al., 2014; Klie et al., 2018). These tools train a machine learning algorithm on a subset of annotations and suggest predicted annotations for new data, producing a hybrid "human-in-the-loop" model (da Silva et al., 2019). Our work here follows in this tradition, seeking effective and efficient methods for collecting reliable new data.

3. Domain Description

For all annotation and experiments in this work, we use transcripts of the 2016 U.S. presidential debates and reaction to the debates on Reddit (Visser et al., 2019). This corpus is appropriate for our task as it includes various rhetorical moves by both politicians and observers in social media. In addition, human annotators have extracted propositions from all dialogues and posts, and identified claim-premise pairs with support and attack relations. Our work focuses on 4,648 propositions that are part of claim-premise pairs with support relations. Approximately half of our data comes directly from debate transcripts, with the remainder coming from social media response. From the transcripts of the debates themselves, approximately 10% of propositions come from moderators while the remainder comes from candidates themselves. The full distributions of speaker affiliations and debate sources are shown in Figure 1.

We are not the first researchers to study this domain. Haddadan et al. (2018) annotated similar presidential debates dating back to 1960, while numerous researchers have studied argumentation on Reddit and similar social media sites (Jo et al., 2018). Datasets have also been developed for similar annotation schemes, like the more syntactically and lexically constrained CommitmentBank (Jiang and de Marneffe, 2019), and for the 2016 U.S. presidential election in particular (Savoy, 2018). Our work, however, is the first to date to examine argumentation frames in this context, at this level of depth, in primarily computational work.

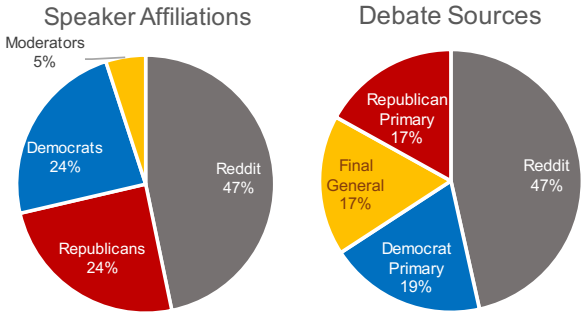

Figure 1: Speaker affiliations and debate sources.

4. Defining Proposition Types

This work does not attempt to cover all of argumentation theory; instead, we focus on four proposition types: normative, desire, future possibility, and reported speech. Using the language from prior work, in our taxonomy *future possibility*, *desire*, and *reported speech* are subtypes of *fact*, while *normative* is close to *policy*. We do not assume that these proposition types are mutually exclusive, choosing to adopt binary annotation for each proposition type. More details and examples are available in the full annotation manuals.

4.1. Normative

A normative proposition is defined as a proposition where the speaker or someone else proposes that a certain situation should be achieved or that an action should be carried out. A normative proposition, under our definition, carries the explicit force of community norms and policies, as opposed to a mere desire or valuation, and includes commands, suggestions, expression of needs, and prohibitive "can't". An example proposition is

> "*the major media outlets* ***should not be the ones*** *dictating who wins the primaries*".

Normative propositions are tightly related to several argumentation schemes. For instance, the argument from consequences scheme proposes that a certain action should (or shouldn't) be carried out because of a potential consequence. Practical reasoning also asserts a normative conclusion in order to achieve a certain goal (Walton et al., 2008). Prior studies have referred to similar normative propositions as "policy" annotations (Park et al., 2015; Egawa et al., 2019).

4.2. Desire

A desire proposition is defined as a proposition that explicitly claims that the speaker or someone else desires to own something, do something, or desires for a certain situation to be achieved. A desire is usually *weaker* than normative propositions and carries no explicit force of proposal or norm. Actively desiring something is also different than merely valuing that thing or asserting a future possibility. An example proposition is:

> "*at the very least for the first debate* ***I'd like to see everyone get a fair shot*** *at expressing themselves*".

In practical reasoning, a normative conclusion is supported by a certain goal to achieve, and this goal is often expressed as another normative proposition or a desire as in

> "*let's have paid family leave, because **I want us to do more to support people who are struggling to balance family and work***."

Prior work has paid little attention to annotating desire propositions. In NLP, the closest work is in subjectivity annotation and the more narrow task of annotating subjectively beneficial events (Somasundaran and Wiebe, 2010; Deng et al., 2013), but these approaches have typically been applied in the context of sentiment analysis; our approach focusing on argument is, to our knowledge, a new contribution in computational linguistics.

4.3. Future Possibility

A future possibility proposition claims a possibility or prediction that something may be the case in the future. These future possibilities are independent of whether the speaker desires the forecast to be true, or believes they *should* be true; the claimed future possibility is just the speaker's own, or someone else's, belief about what the future may hold:

> "*US shooting down a Russian jet **could easily turn ugly***".

Speakers describing their own future plans are also counted as a future possibility. Propositions with future possibilities are often used to support conclusions in the argument from consequences scheme, as in the following example:

> "*Bring us to a 350 ship Navy again, and bring our Air Force back to 2,600 aircraft, because **those are the kind of things that are going to send a clear message around the world***."

An additional scheme, *argument from cause to effect*, also makes use of future possibility as a conclusion, supported by factors that may cause the future event.

4.4. Reported Speech

Our last proposition type is reported speech. A reported speech proposition must convey an explicit or implicit predicate borrowed from a source external to the speaker. We extend the scope of "speech" to belief, thoughts, and questions, in order to capture a wider range of propositional contents borrowed from external sources:

> "***many in the Black Lives Matter movement, and beyond, believe that** overly-aggressive police officers targeting young African Americans is the civil rights issue of our time*".

For each proposition of reported speech, we also annotate text spans that represent the source and the content, and mark the credibility of the source as high, low, or unsure.

Reported speech plays a critical role in discourse; the alignment of a proposition with a third-party source allows for both distancing an author from the claim, and for simultaneously strengthening that claim by appealing to the authority of the original source (Walton et al., 2008). In practice, this is used as a sophisticated rhetorical tool in argument, as a trigger to agree or disagree with the position (Janier and Reed, 2017), to make authority claims (Walton et al., 2008), or even to commit straw man fallacies (Talisse and Aikin, 2006). In the NLP community, a prior study identified authority claims in Wikipedia talk pages (Bender et al., 2011), but the ways of referring to task-oriented norms in these pages are different from general reported speech in argumentation. Park et al. (2015) annotated references (e.g., URLs) in policy-related argumentation, but reported speech was not included as references.

As a methodological note, in the original corpus the pronoun "*I*" has been resolved to the speaker's name in the process of annotating propositions from locutions (e.g., for the sentence "*I believe Americans do have the ability to give their kids a better future*", "*I believe*" has been replaced with "O'MALLEY believes") (Jo et al., 2019). As a result, it is difficult to tell whether the source of a reported speech proposition is indeed the speaker or not. For annotation, we are faithful to the text of each proposition as it is, resulting in many instances of reported speech that can be used for machine learning. Since some of these instances are not reported speech in the original debates, however, instances are not treated as reported speech in our analysis experiments (§6.) if the source and the speaker are identical.

5. Annotation Workflow

The workflow of our annotation process is designed to manage three concurrent problems. First, our annotations require detailed reading of an annotation manual and are difficult to acquire from the minimally trained workers typically used in contexts like crowdsourcing. Second, positive instances are rare (less than 15% of the total dataset for each proposition type), in which case capturing positive instances is challenging but crucial for high inter-annotator agreement and the high quality of annotations. And third, because of the high engagement needed by individual annotators and the lack of positive examples in freely occurring text, the collection and labeling of a dataset sufficiently large to perform quantitative studies and train downstream argument mining classifiers is expensive and logistically challenging.

We solve these problems by leveraging a machine annotator trained on a set of annotations. After we train two human annotators on a subset of data, the remaining corpus is split between them. To expedite annotation, the machine annotator annotates the data first and separates it into a large percentage of instances that are covered by highly reliable machine annotation, mostly of negative instances, and only need quick review by humans, and a remaining small portion that needs to be annotated as usual. To maintain high quality of the final released dataset, as a final step all human annotations are compared with the machine annotations, and discrepancies are resolved by an adjudicator[2].

An overview of this annotation process is shown in full in Figure 2. For each proposition type, our annotation follows a three-stage process. Stage 1 is to train two human annotators. In Stage 2, we train a machine annotator and calibrate it to optimize dataset coverage and accuracy. In Stage 3, the remaining data is annotated by the human and machine annotators in collaboration; final discrepancies are resolved by the adjudicator.

[2]The adjudicator is a co-author of this paper.

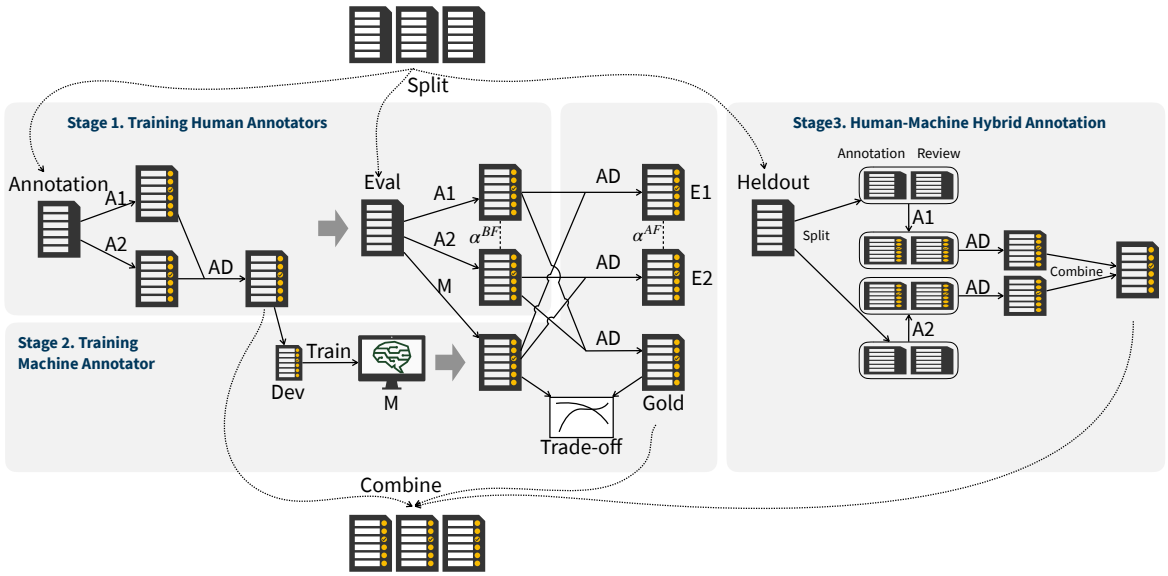

Figure 2: Workflow of annotation process. A1, A2, and AD are two human annotators and an adjudicator, respectively. M is the machine annotator.

Category	Annotation (Dev)	Eval	Heldout Annotation	Heldout Review
Normative	1,497 (924)	400	461	2,290
Future	1,497 (424)	400	433	2,318
Desire	1,497 (424)	400	340	2,411
Rep. Speech	997 (997)	400	541	2,710

Table 1: Statistics of data splits.

5.1. Initial Training for Human Annotators

In this stage, we train two human annotators and evaluate their inter-annotator agreement. We recruited two undergraduate students as annotators; they have no particular experience in argumentation or rhetoric. Approximately 30% of the data (**Annotation**) is used for developing annotation guidelines and training human annotators iteratively over multiple rounds. We then evaluate the final annotation guidelines for reliability on the **Eval** set, approximately 10% of the entire data (Table 1).

Inter-annotator agreement (IAA) was measured using Krippendorff's alpha. We achieve results of $\alpha = 0.67$ for Normative types, $\alpha = 0.59$ for Desire types, $\alpha = 0.66$ for Future types, and $\alpha = 0.71$ for Reported Speech. Despite quite intensive training of human annotators, the main challenge for achieving substantially high IAA is the small number of positive instances; missing a few positive instances greatly affects the IAA score. This motivates our use of the machine annotator as a second annotator.

For reported speech, we also annotated the text spans of sources and contents, and the credibility of the sources. To evaluate the annotators' agreement on sources or contents, we first filtered propositions that both annotators marked as reported speech, and for each proposition, we obtained the longest common sequence of words between two text spans from the annotators. The average number of words that are outside of the common span is 0.5 for sources and 0.2 for contents. Most mismatch comes from articles ("*the experts*" vs. "*experts*") or modifiers ("*President Clinton*" vs. "*Clinton*"). For credibility annotations, the annotators agreed on 85% of the annotations. These results show that the annotations of sources, contents, and credibility are reliable.

5.2. Training Machine Annotator

In this stage, we train a machine annotator and calibrate it to optimize the amount of dataset it covers and annotation accuracy. A subset of the Annotation set is annotated on the final, independently reliable annotation guidelines (**Dev**) and used for training the machine annotator for each proposition type (Table 1). For machine learning feature representation and labeling, we use the single sentence classification model in BERT[3] (Devlin et al., 2018). The input is the full text of a proposition, and the output is the probability that the input proposition is an instance of the corresponding proposition type. Representation is fully automated in a deep neural model that makes extensive use of attention weights and intermediate representations. We used the pretrained uncased, base model with the implementation provided by Hugging Face[4]. The machine annotator's accuracy on the Dev set using 5-fold cross validation is shown in Table 3.

To evaluate how the machine annotator can improve the reliability of annotations, the **Eval** set was also annotated by the machine, and discrepancies between the machine predictions and original human annotations from both annotators were resolved by the adjudicator (E1 and E2 in Figure 2). As expected, the IAA improved significantly from before adjudication (α^{BF}) to after adjudication (α^{AF} in Table 2); the final adjudicated agreement between annotators is between 0.83 and 0.97. The disagreement rate between a human annotator and the machine annotator—annotations that need to be adjudicated—ranges between 2.5% and 13.0%.

[3]We tried logistic regression with extensive hand-crafted features as well, but BERT performed significantly better.

[4]`github.com/huggingface/transformers`

Category	α^{BF}	α^{AF}	DA (A1)	DA (A2)
Normative	0.67	0.97	6.8%	11.0%
Desire	0.59	0.86	2.5%	4.5%
Future	0.66	0.96	4.8%	4.5%
Reported Speech	0.71	0.83	13.0%	13.0%

Table 2: IAA on the Eval set. α^{BF} and α^{AF} are the IAA before and after machine involvement, respectively. "DA (A1)" and "DA (A2)" are the instance-level disagreement rates between the machine and the two human annotators.

Category	Prec	Recl	F1	AUC
Normative	84.1	88.7	86.1	98.1
Desire	100.0	70.0	80.0	95.1
Future	60.0	81.4	62.8	98.2
Reported Speech	44.6	92.9	59.4	96.4

Table 3: Machine performance using 5-fold cross validation.

We next move to questions for developing a hybrid human-machine annotation pipeline. We take advantage of the distribution of classifier output probabilities, finding that the machine annotator has very high AUC scores (Table 3), and that the shape of the probability distributions is well-suited to filtering out instances that are unlikely to contain positive examples of our proposition types. We define a probability threshold k and say that instances with probability of a positive label less than k are *covered* by the model.

We analyzed how k affects the coverage and annotation accuracy on the Eval set. For this analysis, we first created gold standard annotations for the Eval set by the adjudicator resolving disagreements between the annotations of the two human annotators (**Gold** in Figure 2). Then, for each value of k, we replaced the labels of instances whose predicted probability is lower than k with the machine annotator's decisions and measured the IAA and agreement rate between these partially replaced annotations and the Gold set. Figure 3 shows visually the trade-off between this threshold, quantity of data covered, and annotation accuracy as k increases:

- **Dataset coverage (red line):** A large percentage of instances, over half, are clumped together and assigned probabilities of positive labels of approximately $k = 0.2$. After this large group of negative instances comes a steadier growth in coverage between $k = 0.2 - 0.9$.
- **Agreement (Krippendorf's α, blue line; accuracy, yellow line):** This estimates the lower bound of accuracy from human-machine hybrid annotation without final adjudication. Initially, for low values of k, accuracy remains at or approximately 100%, because the machine filters out likely negative instances well. As k grows, overall model accuracy decreases.

This resulting model is a good fit for a hybrid human-machine annotation workflow. The models efficiently filter out negative samples with high coverage at a relatively low value of k, producing a much smaller and more balanced set of candidate propositions for human annotation. Below this

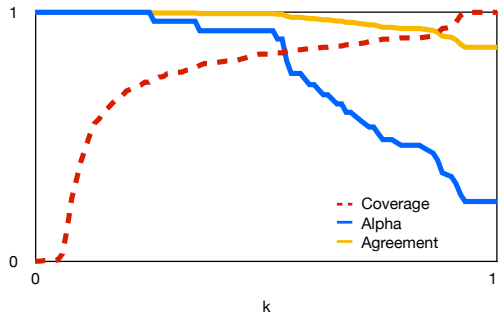

Figure 3: Trade-off between data coverage and annotation accuracy as the threshold of machine-predicted probability k varies. This graph is for Reported Speech, but other proposition types have similar tendencies.

Category	Metric	k	Coverage	α	Agree
Normative	mean	.19	78%	.98	99.0%
Desire	max	.34	81%	.95	99.6%
Future	mean	.39	88%	.95	99.1%
Reported Speech	mean	.35	78%	.96	99.7%

Table 4: Final configurations of the machine annotator.

threshold, instances are assigned negative labels (because $k < 0.5$) and are only subject to very efficient human review; above this threshold, humans are required for a more time-consuming full annotation process. Table 4 shows the hyperparameter selection of mean or max probabilities of the 5-fold classifiers; the tuned threshold k for each proposition type; and the resulting data coverage, α, and agreement rate (accuracy).

5.3. Human-Machine Hybrid Annotation

In the last stage of our workflow, the remaining data (**Heldout**) is split between the two human annotators. Each split is further split into an annotation set and a review set (Table 1); the annotation set is annotated by the human annotator as usual, and the review set is pre-annotated by the machine, and reviewed and corrected by the human annotator. Since human annotators may make mistakes, the annotations of a human annotator for both the annotation and review sets are compared with the machine annotations, and disagreements are resolved by the adjudicator.

Detailed statistics of annotation speed and disagreement rates are listed in Table 5. On average, the review session is three times faster than the annotation session, expediting annotation significantly for a large portion of the data. Both annotators see efficiency boosts of between 39.0% and 85.3%, depending on proposition type, when moving from the full annotation process to review of machine annotations. We observe that the two human annotators have different annotation paces for each proposition type. This situation is common in many annotation tasks where data is split among annotators; although it could potentially result in inconsistent annotations, many annotation studies do not take a further step of quality control. In our task, when all human annotations were compared with the machine annotations, on average 6% of instances had disagreement, which was

Category	A1				A2			
	Annotation	Review	Gain	Agreement	Annotation	Review	Gain	Agreement
Normative	17.3	3.0	82.7%	93.2%	6.4	1.9	70.3%	96.5%
Desire	8.3	3.4	59.0%	96.5%	4.9	1.8	63.3%	95.0%
Future	10.4	5.3	49.0%	93.0%	10.9	1.6	85.3%	99.0%
Reported Speech	10.6	6.5	39.0%	86.6%	22.4	7.1	67.4%	91.2%

Table 5: Annotation speed (sec/proposition) and efficiency gain moving from full annotation to review of machine labels, and instance-level agreement rates between single human and machine annotation on the Heldout set.

Normative	Desire	Future	Reported Speech	Total
602 (13%)	147 (3%)	453 (10%)	242 (5%)	4,648

Table 6: The number of positive instances and their proportion for each proposition type for the entire data.

resolved by the adjudicator (Table 5). This emphasizes the value of our approach, using a machine annotator to double check human annotations and resolve potentially incorrect annotations with a small effort of adjudication. The prevalence of each proposition type for the entire released dataset is shown in Table 6. Labels are not exclusive or conditioned on each other; in total, 30% of the final dataset contains at least one positive annotation, and most other positions describe judgments and facts.

6. Analysis of U.S. Presidential Debates

Our annotations readily allow us to conduct some interesting analyses of the 2016 U.S. presidential debates. First, different speakers in the debates use different rhetorical strategies, and our proposition types shed light on how the strategies differ in terms of the kinds of statements made by the speakers. Next, we analyze varying types of claims made in the debates and what types of premises are commonly used to support those claims.

6.1. Use of Proposition Types by Main Speakers

Across individual speakers: As representative examples of how our annotations can be used to evaluate language in use, we first chose five main speakers to examine how they differ in their use of proposition types: Donald Trump, Hillary Clinton, Bernie Sanders, Anderson Cooper, and Reddit users (as an aggregated group). Trump and Clinton were the nominees of the Republican and Democratic Parties, while Sanders was a competitive rival of Clinton. Cooper was a main moderator of the debates. For each of these speakers, we calculated the proportion of each proposition type and then normalized these proportions to z-scores.

As shown in Figure 4, these five exemplar speakers use proposition types differently (their distributions of the types are significantly different with $p <$ 1e-5 for a χ^2 test). When compared to Trump, the Democratic candidates make much greater use of normative language. In particular, language from the two Democratic candidates uses normative propositions and expresses desires a lot more than Trump, often to make the case for specific policies based on normative

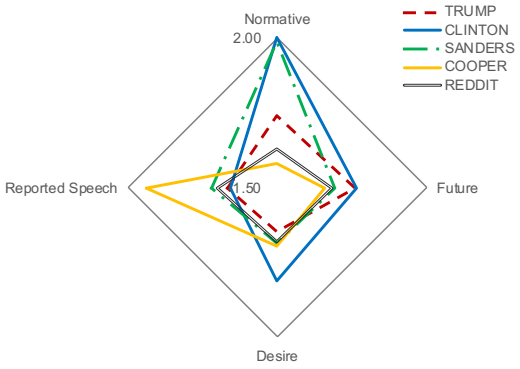

Figure 4: Use of proposition types by five main speakers, normalized to z-scores.

values. Clinton makes the most use of normative language, while Sanders uses future possibilities less and reported speech slightly more. Again, though, the major differentiator is in normative language, where he mirrors Clinton.

> Clinton: "*We also, though, need to have a tax system that rewards work and not just financial transactions*"
>
> Sanders: "*War should be the last resort that we have got to exercise diplomacy*"

These normative judgments are not absent entirely from Trump's language, but they are less prevalent. Among moderators, Cooper uses significantly more reported speech and less normative claims and future possibilities than candidates or online commenters, which matches his role as a moderator in contemporary politics. While early debates in the television era leaned on questions from moderators that were "unreflective of the issues salient to the public" (Jackson-Beeck and Meadow, 1979), moderators today view themselves as serving a "gatekeeping" function that is nevertheless representative of a curated version of engaging questions that are relevant to the public interest, expressed through reported speech (Turcotte, 2015). Lastly, Reddit users make use of less rhetorical structure than candidates of either party or moderators, instead focusing more on past/current events, facts, and more straightforward rhetoric. This is reflected in their lower use of normative propositions, future possibilities, and desire compared to the candidates.

Across affiliations: Next, we examined whether there is a significant difference in use of propositions types among Republican candidates, Democratic candidates, and Reddit users. We split propositions into the three groups (excluding

	Normative	Desire	Reported Speech
Dem vs. Rep	+++	+	
Reddit vs. Dem	−−−	−	++
Reddit vs. Rep	−−−		+

Table 7: Comparison of proposition types used by Republicans, Democrats, and Reddit users. +/− represents whether the group on the left uses a higher/lower proportion of the proposition type than the group on the right, and the numbers of +/− indicate significance levels (one: $p < .05$, two: $p < .01$, three: $p < .001$). There was no significant difference in use of future possibilities among the groups.

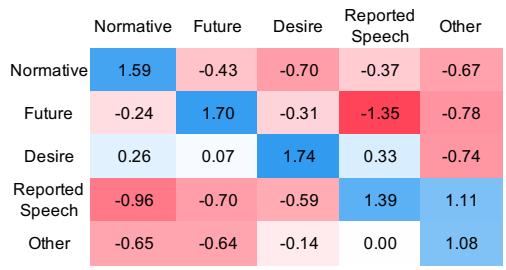

Figure 5: Normalized z-scores of correlations between proposition types in claim-premise pairs. Rows are claim types and columns are premise types.

moderators) and tested for differences in proportion of each proposition type across groups, using χ^2 tests.

As shown in Table 7, Democratic candidates as a whole continue the trend we observe in individual speakers. They use more normative propositions and desire expressions than Republican candidates, and this result across groups is highly significant. However, they had no significant difference in use of reported speech and future possibilities. Reddit users make less use of argumentation proposition types in general: they use less normative language than the candidates and express less desire than Republican candidates. However, they use reported speech often, partly because their discussions occurred after the debates had occurred. As a result, these texts often refer back to speech from the debates themselves and the reported speech of the candidates.

6.2. Proposition Types in Claim-Premise Pairs

The propositions in our work are drawn from claim-premise pairs, as annotated in the original corpus. As such, we are able to merge our annotations with this pre-existing structure for deeper analysis. We do so first by examining the correlations of the proposition types between claims made in the debates and their supporting premises. We computed the correlations between proposition types in claim-premise pairs as follows. First, since a few propositions have more than one proposition type, we chose the main type in the importance order of normative, desire, reported speech, and future possibility. Propositions that do not belong to any of these types are classified as other. For each type of claims, we calculated the distribution over proposition types for their premises and normalized again to z-scores (Figure 5).

Each proposition type has different degrees of correlations with other proposition types. Naturally, proposition types often match between claims and premises—the appearance of a particular proposition type in a premise conditioned on that type appearing in a claim is high (the diagonal of the table). We see many instances of normative claims supported by a normative premise, constituting practical reasoning:

> "*We need to control our border, because it's our responsibility to pick and choose who comes in.*"

Similarly, many claims of future possibility are supported by the same type of premise, constituting an argument from cause to effect:

> "*Families' hearts are going to be broken, because their kids won't be able to get a job in the 21st Century.*"

On the other hand, certain pairings are deeply unnatural and rarely-occurring in natural text. Pairs comprised of future-looking claims based on premises from reported speech, for instance, are the least likely pairing in our dataset. The correlation analysis supports our belief that proposition types can be useful information for studying argument.

7. Conclusion

This analysis has application to tasks like argument generation, where correlation information may inform systems of what kind of premise should likely follow a certain type of claim in natural speech, allowing parameterization beyond mere topic and into NLG that controls for style and structure, a goal of recent work (Prabhumoye et al., 2018). For argumentation scheme annotation, high correlations between proposition types imply that the proposition types may reflect different argumentation schemes, and may provide a structured crosswalk between argumentation theory, which is often nuanced and resists quantification at scale, and NLP advances that are often limited to labeling tasks.

Through the introduction of this new corpus of the U.S. 2016 presidential debates and commentary, annotated with four proposition types that capture nuanced building blocks of argumentation schemes, we hope to advance the state of the art in argument mining. For effective annotation, we presented a human-machine hybrid annotation protocol that allows for efficient and reliable annotation for difficult annotation tasks involving complex reasoning and rare occurrences of positive instances; we believe this methodology is replicable in the identification, annotation, and study of sociolinguistic or argument features more broadly that appear rarely. Today's machine learning systems struggle with such skewed distributions in a fully automated context, but we demonstrated that both the speed and inter-annotator reliability of these annotations can be enhanced with a hybrid approach that makes targeted, selective use of machine learning methods. Future research should test whether the distributional properties that make this approach effective in our domain, like high recall and near-100% precision in low-probability negative instances, are part of a more general pattern in annotation of rare linguistic phenomena in text and speech.

8. Acknowledgements

This research is supported by the Kwanjeong Educational Foundation and in the UK under EPSRC grant EP/N014871/1.

9. Bibliographical References

Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., and Ghobadi, M. (2017). What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

Alonso, O., Marshall, C. C., and Najork, M. (2015). Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 101–110. ACM.

Bender, E. M., Ostendorf, M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., and Zhang, B. (2011). Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57.

Carrell, D. S., Cronkite, D. J., Malin, B. A., Aberdeen, J. S., and Hirschman, L. (2016). Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification. *Methods of information in medicine*, 55(04):356–364.

Cocarascu, O. and Toni, F. (2018). Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, 44(4):833–858.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

da Silva, T. L. C., Magalhães, R. P., de Macêdo, J. A., Araújo, D., Araújo, N., de Melo, V., Olímpio, P., Rego, P., and Neto, A. V. L. (2019). Improving named entity recognition using deep learning with human in the loop. In *EDBT*, pages 594–597.

Deng, L., Choi, Y., and Wiebe, J. (2013). Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the Association for Computational Linguistics*, pages 120–125.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, October.

Duthie, R., Budzynska, K., and Reed, C. (2016). Mining ethos in political debate. In *COMMA*, pages 299–310.

Egawa, R., Morio, G., and Fujita, K. (2019). Annotating and Analyzing Semantic Role of Elementary Units and Relations in Online Persuasive Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 422–428, Florence, Italy, July. Association for Computational Linguistics.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Frau, J., Teruel, M., Alemany, L. A., and Villata, S. (2019). Different flavors of attention networks for argument mining. In *Proceedings of FLAIRS*.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Haddadan, S., Cabrio, E., and Villata, S. (2018). Annotation of argument components in political debates data. In *Proceedings of the Workshop on Annotation in Digital Humanities*.

Hoffman, E. R., McDonald, D. W., and Zachry, M. (2017). Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):52.

Hollihan, T. A. and Baaske, K. T. (2015). *Arguments and Arguing: The Products and Process of Human Decision Making, Third Edition*. Waveland Press.

Imran, M., Mitra, P., and Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 1638–1643. European Language Resources Association (ELRA).

Jackson-Beeck, M. and Meadow, R. G. (1979). The triple agenda of presidential debates. *Public Opinion Quarterly*, 43(2):173–180.

Janier, M. and Reed, C. (2017). I didn't say that! Uses of SAY in mediation discourse:. *Discourse Studies*, 19(6):619–647, August.

Janier, M. and Saint-Dizier, P. (2019). *Argument Mining*. Linguistic Foundations. Wiley, 1 edition, October.

Jiang, N. and de Marneffe, M.-C. (2019). Evaluating bert for natural language inference: A case study on the commitmentbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6088–6093.

Jo, Y., Poddar, S., Jeon, B., Shen, Q., Rose, C., and Neubig, G. (2018). Attentive interaction model: Modeling changes in view in argumentation. 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 103–116.

Jo, Y., Visser, J., Reed, C., and Hovy, E. (2019). A Cascade Model for Proposition Extraction in Argumentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 11–24, Florence, Italy, August. Association for Computational Linguistics.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Lasersohn, P. (2009). Relative truth, speaker commitment, and control of implicit arguments. *Synthese*, 166(2):359–374.

Lawrence, J. and Reed, C. (2019). Argument Mining: A Survey. *Computational Linguistics*, 0(0):1–54, August.

Lawrence, J., Visser, J., and Reed, C. (2019). An Online Annotation Assistant for Argument Schemes. In *The 13th Linguistic Annotation Workshop*, pages 1–8, July.

Leidner, J. L. and Plachouras, V. (2017). Ethical by design: ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40.

Lindahl, A., Borin, L., and Rouces, J. (2019). Towards Assessing Argumentation Annotation - A First Step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence, Italy, August. Association for Computational Linguistics.

Lippi, M. and Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25, March.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

Lugini, L. and Litman, D. (2018). Argument component classification for classroom discussions. *Proceedings of Empirical Methods in Natural Language Processing*, page 57.

Mayfield, E., Adamson, D., and Rosé, C. P. (2013). Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 104–113.

Naderi, N. and Hirst, G. (2015). Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems*, pages 16–25. Springer.

Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *AAAI Conference on Artificial Intelligence*.

Niculae, V., Park, J., and Cardie, C. (2017). Argument mining with structured svms and rnns. In *Proceedings of the Association for Computational Linguistics*, pages 985–995.

Nussbaum, E. M. (2011). Argumentation, Dialogue Theory, and Probability Modeling: Alternative Frameworks for Argumentation Research in Education. *Educational Psychologist*, 46(2):84–106, April.

Park, J. and Cardie, C. (2018). A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Park, J., Blake, C., and Cardie, C. (2015). Toward Machine-assisted Participation in eRulemaking: An Argumentation Model of Evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210, New York, NY, USA. ACM.

Persing, I. and Ng, V. (2016). End-to-end argumentation mining in student essays. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1384–1394.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., and Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Potter, J. (1996). *Representing reality: Discourse, rhetoric and social construction*. Sage.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. In *Proceedings of the Association for Computational Linguistics*, pages 866–876.

Savoy, J. (2018). Trump's and clinton's style and rhetoric during the 2016 presidential election. *Journal of Quantitative Linguistics*, 25(2):168–189.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

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

Song, Y., Deane, P., and Beigman Klebanov, B. (2017). Toward the Automated Scoring of Written Arguments: Developing an Innovative Approach for Annotation. *ETS Research Report . . .* , 152(3):157, April.

Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Talisse, R. and Aikin, S. F. (2006). Two Forms of the Straw Man. *Argumentation*, 20(3):345–352, September.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. 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 809–819.

Turcotte, J. (2015). The news norms and values of presidential debate agendas: An analysis of format and moderator influence on question content. *Mass Communication and Society*, 18(3):239–258.

Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2019). Argumentation in the 2016 US Presidential Elections . *Language Resources and Evaluation*, pages 1–35, January.

Wagemans, J. H. M. (2016). Constructing a Periodic Table of Arguments. In *OSSA Conference Archive*, pages 1–13, December.

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., and

King, J. (2012). A corpus for research on deliberation and debate. In *LREC*, pages 812–817. Istanbul.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press, August.

Wentzel, D., Tomczak, T., and Herrmann, A. (2010). The moderating effect of manipulative intent and cognitive resources on the evaluation of narrative ads. *Psychology & Marketing*, 27(5):510–530.

Woods, B., Adamson, D., Miel, S., and Mayfield, E. (2017). Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2071–2080. ACM.

Yimam, S. M., Biemann, C., Eckart de Castilho, R., and Gurevych, I. (2014). Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland, June. Association for Computational Linguistics.