# GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates


Vibhor Agarwal
Department of Computer Science
University of Surrey
Guildford, Surrey, United Kingdom
v.agarwal@surrey.ac.uk

Sagar Joglekar
Department of Informatics
King's College London
London, United Kingdom
sagar.joglekar@kcl.ac.uk

Anthony P. Young
Department of Informatics
King's College London
London, United Kingdom
peter.young@kcl.ac.uk

Nishanth Sastry
Department of Computer Science
University of Surrey
Guildford, Surrey, United Kingdom
n.sastry@surrey.ac.uk



## ABSTRACT

Online forums that allow participatory engagement between users have been transformative for public discussion of important issues. However, debates on such forums can sometimes escalate into full blown exchanges of hate or misinformation. An important tool in understanding and tackling such problems is to be able to infer the argumentative relation of whether a reply is supporting or attacking the post it is replying to. This so called polarity prediction task is difficult because replies may be based on external context beyond a post and the reply whose polarity is being predicted. We propose GraphNLI, a novel graph-based deep learning architecture that uses graph walk techniques to capture the wider context of a discussion thread in a principled fashion. Specifically, we propose methods to perform root-seeking graph walks that start from a post and captures its surrounding context to generate additional embeddings for the post. We then use these embeddings to predict the polarity relation between a reply and the post it is replying to. We evaluate the performance of our models on a curated debate dataset from Kialo, an online debating platform. Our model outperforms relevant baselines, including S-BERT, with an overall accuracy of 83%.



## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Machine learning algorithms**; **Model development and analysis**; *Information extraction*; • **Information systems** → **World Wide Web**.

## KEYWORDS

Online debates, argument mining, polarity prediction, Kialo




*WWW '22, April 25–29, 2022, Virtual Event, Lyon, France*



ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512144

**ACM Reference Format:**
Vibhor Agarwal, Sagar Joglekar, Anthony P. Young, and Nishanth Sastry. 2022. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512144


## 1 INTRODUCTION

The Internet has enabled people to participate in sharing their views, often as comments or posts, about many topics online. Online debates can sometimes become large and acrimonious, with some escalating into full blown exchanges of hate and misinformation. As many of these debates concern topics of societal importance, it is crucial to be able to model these debates accurately and at scale, so that we can better understand and control phenomena such as the spread of hate [15, 28], fake news [4, 24], how best to moderate political polarisation [5], and how to break echo chambers by linking appropriate users of opposing views [22].

An important task in modelling online debates is to be able to predict whether the reply of one comment to another is **attacking** (disagreeing) or **supporting** the post it is replying to. This is called the **polarity** of the reply [13]. The ability to accurately predict the polarity of replies in an online debate can allow us to measure properties of the debate, such as how "controversial" a discussion is, *e.g.* by counting the number of supporting *vs.* attacking replies [10]. Perhaps more importantly, if the polarity is known, we can then use techniques from argumentation theory, a branch of artificial intelligence concerned with the formal representation and resolution of disagreements [37], to formally compute which arguments have been attacked and rebutted, and which ones stand unrebutted or are further justified by additional supporting replies.

One obvious approach to predicting the polarity of replies in any debate is to apply natural language processing (NLP) techniques. Such models typically take as input the natural language text of a comment $b$ and the comment $a$ that it is replying to, and output a predicted polarity for whether $b$ is attacking or supporting $a$ (*e.g.* [11, 17]). However, one shortcoming of such an approach is that it risks loosing crucial information by considering the comments $a$ and $b$ in isolation from the rest of the debate.

We call the posts $a$, $b$ as the *local* context for the polarity prediction task and the whole discussion thread in which these comments are embedded as the *global* context. This paper asks and answers the question: *Can we improve performance on the polarity prediction task by incorporating additional context beyond the posts a, b?*

Typically, discussion threads can be seen as a tree, starting with an original post which forms the root of the tree and each reply such as $b$ creating a directed edge in the tree, from $b$ to its parent node $a$ that it is replying to. We hypothesise that nodes near $a$ and $b$, *e.g.* their children, ancestors and siblings in the discussion tree, contain additional information that may help understand whether $b$ is attacking $a$ or supporting it. For example, if other siblings of $b$ (*i.e.* children of $a$ other than $b$) are also attacking $a$, then it may be more likely that $b$ is also an attacking reply. *Our key idea is to use graph walk techniques to discover and utilise neighbouring context in a principled fashion.* Our contributions are as follows:

(1) We compare and contrast several NLP models, including Sentence-BERT [38], to establish a baseline for the polarity prediction task. We use data from Kialo, an online debating platform[1]; this data is in the form of discussion trees where the nodes are arguments submitted to the debate[2] and the edges denote which arguments reply to which other arguments, as will be explained in Section 4.1.
(2) We propose graph walk techniques that sample discussion trees with the aim of capturing parts of the global context of the online debates, and input the additional nodes sampled into deep learning model along with the local context of the replying argument and the argument being replied to. We find that a weighted root-seeking graph walk works the best in capturing the wider context of the debates on Kialo.
(3) We present and evaluate **GraphNLI** – a novel graph-based deep learning architecture to predict the polarity of replies. Our model outperforms baselines including S-BERT, achieving an accuracy of 83%. We provide an open source implementation of the model available for public usage[3].
(4) We systematically investigate through ablation studies what features can be helpful in capturing the wider context for the polarity prediction task and show that upstream (or earlier) text, *i.e.*, the parent and other ancestor nodes, help the model more than siblings and children replies. Moreover, we find that the importance of ancestor nodes decreases as their distance from the given node increases.

The rest of this paper is structured as follows. Sec. 2 provides an overview of the polarity prediction task in the context of argumentation theory. In Sec. 3, we discuss the details of Kialo dataset. We next consider different ways of incorporating the wider context of discussions in Sec. 4, and present the GraphNLI architecture. We then evaluate the model in Sec. 5, comparing its performance relative to a range of baseline classifiers and conducting an ablation study to better understand which features are important. In Sec. 6, we summarize our results, and outline possible future work.

[1]See https://www.kialo.com/, last accessed 17 October 2021.
[2]We will use the terms "node", "post", "comment" and "argument" interchangeably.
[3]The model code is available at https://github.com/socsys/GraphNLI and the dataset is available at https://tinyurl.com/kialo-debates-dataset.

## 2 BACKGROUND AND RELATED WORK

**Argumentation theory** is a branch of AI that is concerned with the transparent and rational resolution of disagreements (e.g. [37]). Many formal models of argumentation have been devised and studied, starting with **abstract argumentation theory** [6, 20, 40], where the arguments of interest are elements of a set $A$, and a not-necessarily-symmetric binary relation $R \subseteq A \times A$ - the **attack relation** - represents when two arguments disagree, i.e. $(b, a) \in R$ denotes that argument $b$ disagrees with argument $a$. The directed graph (digraph) $\langle A, R \rangle$ is called an **abstract argumentation framework** (AF); this abstracts away from the structure of the arguments and the nature of their disagreements. Resolving these disagreements formally amounts to identifying subsets of arguments $S \subseteq A$ that satisfy various normative properties. For example, the property of **conflict-freeness**, $(S \times S) \cap R = \emptyset$, formalises the idea of self-consistency because winning arguments should not attack each other.

Of course, arguments can agree as well as disagree. **Bipolar argumentation frameworks** (BAFs) enrich AFs with a **support relation** that models when arguments agree [13]. Formally, a BAF is a digraph $\langle A, R_{att}, R_{sup} \rangle$ where $R_{att}, R_{sup} \subseteq A \times A$ are two disjoint binary relations respectively denoting attack and support between arguments - every edge of a BAF is either attacking or supporting but never both. To determine the winning arguments, there are various ways to transform a BAF into an AF by "absorbing" supports into attacks (e.g. [14]) depending on how supports are interpreted. Following this literature, given an edge in a BAF, we call the status of whether this edge is attacking or supporting its **polarity**.

Both AFs and BAFs are theoretically interesting and mathematically elegant, and offer a normative and perhaps intuitive way of identifying winning arguments in the presence of agreement and disagreement [13, 20, 36]. However, in order to apply these models to reason about online debates, we need to map the concepts identified in these models to their real-life counterparts. **Argument mining** (e.g. [12, 31, 32]) is the application of NLP tools to extract arguments and identify their relationships from raw text. Example tasks involve identifying when Tweets from Twitter are well-defined arguments instead of insults, single URLs or pictures (e.g. [9]), identifying the claims, their reasons and relationships between claims from clinical trials to inform medical decision making (e.g. [34]), or detecting fallacies from the transcripts of the United States Presidential Debates [39]. Online debates, such as those on Reddit or Twitter for example, can be converted into the argumentation framework by first defining nodes or arguments corresponding to each well-defined logical claim, and drawing signed edges representing supports or attacks between them. For instance, when a comment $b$ replies to another comment $a$, we may have two nodes $a, b$ linked by an edge representing a support or an attack depending on the polarity relation between them. The result, therefore, is to formally represent an online debate as a BAF.

Many other analyses can be performed once a debate has been represented as a BAF. For instance, based on the polarities of the edges, we may calculate which arguments are justified and which arguments have been rebutted. This could potentially be used to

present only the justified arguments as summary to a reader. Previous work has also looked at how the conclusions of a logical reader can change depending on which parts of a debate they read, thus underscoring the dangers of sampling only parts of a large online debate [41–43]. Other work has shown how the location of justified arguments can be significantly influenced by whether the debate is acrimonious or supporting [10]. BAFs are thus useful representational tools for modelling online debates, allowing the application of both argument-theoretic and graph-theoretic ideas to gain insights about online discussions.

The polarity prediction task for online debates has been addressed in the argument mining literature.[4] An early example is [11], which applied textual entailment [8, 18, 29] to predict the polarity of replies on the now-defunct Debatepedia dataset, with a test accuracy of 67%. In [17], long-short-term memory networks were used to classify polarity, achieving 89% accuracy.[5] A more recent overview of the polarity prediction task [16] has provided context-independent baselines of neural network models using a range of learning representations and architectures, and have found an averaged performance of 51% to 55% of these different neural networks across such contexts; these contexts involve online debates on controversial topics such as abortion and gun rights, persuasive essays, and presidential debates.

In all of the above-mentioned approaches, the inputs to the model are the texts of the replying argument and the argument being replied to, often represented by some appropriate word embedding. Arguably, this is the least amount of information one must input into the model to predict the polarity of the reply. What has not yet been considered is whether it is helpful to input more information. For example, if argument $c$ replies to argument $b$, and $b$ replies to argument $a$, and we would like to predict the polarity relation between $c$ and $b$, then is it useful to design a model that accepts as input the texts of $c$, $b$ *and* $a$? How about if we randomly sample additional "nearby" comments? What if we incorporate graph-theoretic features into the input such as in-degree? To the best of our knowledge, these questions have not yet been addressed in the argument mining literature. We thus seek to investigate these questions by measuring whether the incorporation of such features can improve the polarity prediction accuracy.

## 3 KIALO DATASET

Kialo is an online debating platform that helps people "engage in thoughtful discussion, understand different points of view, and help with collaborative decision-making".[6] In this study, we use data from discussions hosted on the Kialo debating platform as used by [10]. Figure 1 illustrates an example Kialo discussion. In a Kialo debate, users submit **claims**. The starting claim of a debate is its **thesis**. To start a discussion in Kialo, the user creates a thesis along with a tag that indexes the discussion by indicating the content of the discussion. A thesis can have many tags, which increases its visibility to the users. Then users comment on the discussions of their choice, which as the discussion develops, takes the shape of a directed tree — this is because each non-thesis claim replies to exactly one other claim. The dataset contains 1,560 discussion threads, and is the most complete snapshot of Kialo, as of 28 January 2020.[7] Each discussion thread has data about the tree structure, votes on each argument's impact on the debate it has been submitted to, and the arguments' texts. Further, each reply between arguments is clearly labelled as attacking or supporting. There is also supplemental metadata such as the time of posting, the time of editing, and the author metadata. All discussions crawled from Kialo thus have a tree structure with a root node that represents the main thesis and each other node is a reply to its parent which either supports or attacks the parent. On each topic, there is a reasonable amount of debate, with a mean of 204 and a median of 68 arguments (standard deviation 463). Kialo debates are typically balanced, with the vast majority of discussion trees having between 40% to 60% of its edges as supporting edges, with the rest being attacking edges.

Due to Kialo's strict moderation policy, each piece of text submitted to a debate is a self-contained argument that has clear claim backed by reasoning. Thus, each post in Kialo can be taken as a node and directed edges can be drawn based on which post is replying to which other post. The polarity prediction task is to decide whether these edges are attacking or supporting.

## 4 THE POLARITY PREDICTION MODEL

As stated in Section 2, polarity prediction is an important task in argument mining. It aims to identify the argumentative relations of attack and support between natural language arguments by classifying pairs of text accordingly, where in our case such texts are comments submitted to online debates, and one text is replying to another text. Various deep learning models have been used in the literature to perform the polarity prediction task. However, they usually consider a pair of texts for the polarity prediction. In this section, we propose a novel graph-based deep learning architecture that not only considers as input the pair of texts, but also systematically captures the context of nearby comments via graph walks.

### 4.1 Representing Debates as Discussion Trees

For every online debate $D$ on Kialo, we construct a tree structure, where a node represents a post and a *directed* edge from a node to its parent, the post it is replying to. Each such edge has an associated label, *support* or *attack*, depending upon whether the comment is respectively for or against its parent comment. The root node of this discussion tree represents the thesis (topic) of a debate.

### 4.2 GraphNLI Architecture

We propose a novel graph-based deep learning architecture which captures both the local and the global context of the online debates through graph-based walks. The GraphNLI architecture is shown in Figure 2, and will be explained in Section 4.2.2.

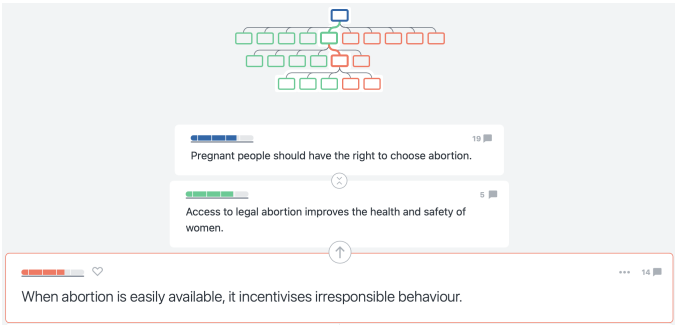
*4.2.1* ***Capturing Global Context through Graph Walks.*** Our GraphNLI model captures the global context of online debates through graph-based walks. A walk is defined as a sequence of

[4] See [12] for reviews of the polarity prediction task in other settings, such as persuasive essays or political debates.

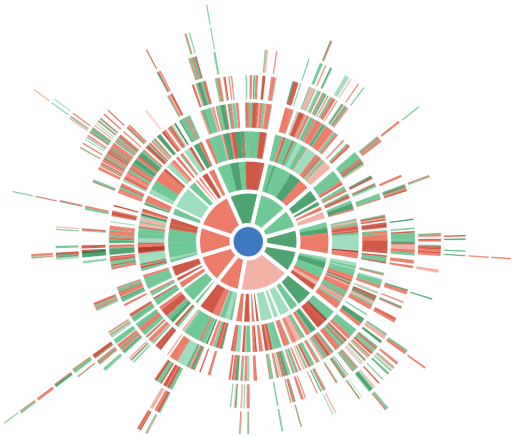
[5] Although this shows better results on the polarity prediction task than what we report here, neither their data nor their framework were available for benchmarking.

[6] Quoted from https://www.kialo.com/about, last accessed 21 Oct 2021.

[7] We collected data continuously until a change in the Kialo backend in January 2020 which made our crawler obsolete. This was sufficient data to test GraphNLI, so no more data was collected.




**(a) An example of the arguments made in a Kialo debate. The thesis is, "*Pregnant people should have the right to choose abortion*". A supporting reply is, "*Access to legal abortion improves the health and safety of women*". An attacking reply to this support is, "*When abortion is easily available, it incentivises irresponsible behaviour*".**



**(b) Visualization of this debate's tree**

**Figure 1: Example of a Kialo discussion. Every debate on Kialo would start with a thesis, which in this example is "Pregnant people should have the right to choose abortion". This thesis can be either supported (shown in green) or attacked (shown in red), and the exchange can go on at multiple levels as seen in Figure 1b. Both figures are taken from https://www.kialo.com/when-abortion-is-easily-available-it-incentivises-irresponsible-behaviour-5637.1340?path=5637.0~5637.1_5637.10160-5637.1340&active=_5637.3912, last accessed 25 Jan 2022.**

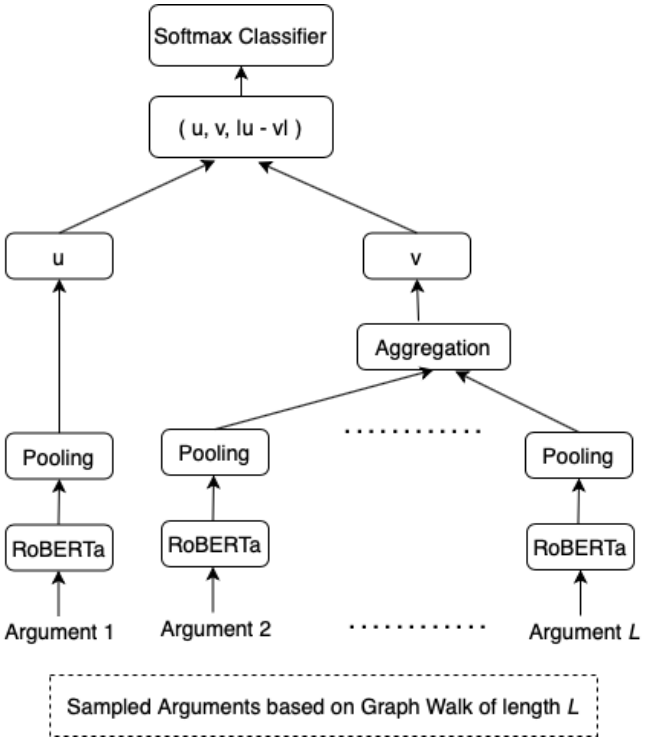



**Figure 2: GraphNLI Architecture**

nodes traversed from a given node in a tree. *We choose to ignore the direction of the edges in these walks*, to enable capturing context such as siblings of the reply node (reachable by going up to the parent node and then back down again to the sibling). Walk length $L \in \mathbb{N}$ is the maximum allowed length after which the graph walk terminates. We propose the following strategies for graph walks in discussion trees to capture the global and neighbouring context of a given argument (node).

**Weighted Root-seeking Graph Walk**: In a discussion tree, a root-seeking graph walk is a walk starting from a given node *up* towards the root of the discussion tree. In a tree, there is exactly one path that goes upwards to the root of the tree from a given node. Depending on the walk length $L$, the graph walk terminates irrespective of whether or not it reaches the root node. We experiment with different walk lengths and find that $L = 5$ gives the best results (see Section 5). The graph walk is *weighted*, which means that the contributions of the ancestor nodes in the node's embedding vector are discounted by a factor of $\gamma^k$, where $k$ is the distance of the node from the given node and $\gamma$ is the *discount parameter*. Therefore, the highest weight is given to the node's parent, then a discounted weight to the parent's parent, and so on. This means the closer the node is to the starting node, the higher its weight will be, and the more it will contribute as an input towards predicting the polarity.

*Example 4.1.* Suppose we have a section of a discussion tree with arguments $a_0$, $a_1$, $a_2$, $a_3$, $a_4$ and $a_5$, where the edges are $(a_k, a_{k+1})$ for $0 \leq k \leq 4$, and $a_{k+1}$ is the parent of $a_k$. As $L = 5$, a weighted root-seeking graph walk from $a_0$ will sample arguments $a_1$ to $a_5$. As for their weights, suppose we use a discount parameter $\gamma = 0.5$, we compute the weight of $a_k$ as $\gamma^k$. Therefore, the weight of $a_0$ is 1, the weight of $a_1$ is 0.5, the weight of $a_2$ is 0.25, and so on.

*Example 4.2.* Suppose we have a section of a discussion tree with arguments $a_0$, $a_1$, $a_2$, where the edges are $(a_k, a_{k+1})$ for $0 \leq k \leq 2$, and $a_{k+1}$ is the parent of $a_k$. Further, assume $a_2$ is the root of the tree. As $L = 5$, a weighted root-seeking graph walk from $a_0$ will

sample arguments $a_1$ and $a_2$, and no more as $a_2$ is the root of the tree itself.

**Biased Root-seeking Random Walk**: This is a random walk starting from a given node in the tree and is biased towards the root. From a given node, each immediate neighbor is assigned a probability of being selected by the random walk. Since the discussion thread is a tree, each node has one parent and zero or more children. To bias the walk towards the root (or thesis node), we assign a probability $p > 0.5$ to the random walk selecting the parent of the given node as the next step in the walk. Recall that we ignore the direction of edges in these walks, allowing the walk to traverse downwards as well as upwards in the discussion tree. Therefore, the remaining probability, $1 - p < 0.5$, is divided equally among all the children of the given node. The walk length $L \in \mathbb{N}$ determines the maximum length of this random walk, that is, the number of nodes to be visited from the given node until the random walk terminates. After experimentation, we find $L = 4$ to be optimal (see Section 5).

The random walk is a different way of sampling the neighbouring nodes as a means of incorporating the surrounding global context for predicting the polarity. Note that there is no guarantee that the parent node will be visited in the random walk. By choosing the probability $p$, we can directly affect the probability of visiting the parent. Empirically, we find that $p = 0.75$ gives the best results. Even if the parent is not visited, there is likely to be information in the surrounding nodes that still helps predict the polarity of the parent-child relationship. For example, if the majority of children nodes replying to the parent are attacking (perhaps because the parent post makes a controversial or wrong statement), knowing the sibling context may help predict the polarity of a reply.

*Example 4.3.* Suppose we have a section of a discussion tree with nodes $a_0, a_1, a_2, a_3$ and $a_4$, with edges $(a_1, a_0)$, $(a_2, a_1)$, $(a_3, a_1)$ and $(a_4, a_1)$, such that for $(a_i, a_j)$, $a_j$ is the parent of $a_i$. Let $p = 0.75$, then $1 - p = 0.25$. A random walk starting at $a_1$ will have probability 0.75 moving to $a_0$ next. Similarly, starting from $a_1$, there is a probability $\frac{0.25}{3} = \frac{1}{12}$ of moving to any of $a_2$, $a_3$, or $a_4$ next.

Also, the same node can be visited multiple times in a random walk, especially when the graph walk chooses one of the children nodes:

*Example 4.4.* Suppose we have a section of a discussion tree with nodes $a_0, a_1, a_2, a_3$ and $a_4$, with edges $(a_1, a_0)$, $(a_2, a_1)$, $(a_3, a_1)$ and $(a_4, a_2)$, such that for $(a_i, a_j)$, $a_j$ is the parent of $a_i$. Let $p = 0.75$, then $1 - p = 0.25$. Assume that a random walk, starting at $a_1$, moves to $a_2$ with probability $\frac{0.25}{2} = \frac{1}{8}$. Now from $a_2$, the random walk can either move to $a_1$ *again* with probability 0.75 or to $a_4$ with probability 0.25.

At the end of a graph walk for each node, we obtain at most $L$ arguments which are either parents or neighbouring responses of a given node. We input these sets of arguments into our GraphNLI model as described in the next section.

*4.2.2* ***Model Overview.*** Our model is inspired by S-BERT [38]; its architecture is shown in Figure 2. Firstly, each of the $L$ arguments sampled by the graph walk or the random walk is input into the RoBERTa model [33] to get their corresponding embeddings and then, a mean-pooling operation, that is, calculating the mean of all the output vectors, is applied to derive a fixed-sized sentence embedding for each argument. The starting node in a graph walk is a *point-of-interest* (PoI) node. Let $u$ denote the sentence embedding corresponding to the PoI node. Let $v$ denote the aggregated embedding from its neighbors (including parent) sampled by the graph walk starting from the PoI node. These $u$ and $v$ embeddings together are used in GraphNLI model to predict the polarity. We experiment with three aggregation strategies: *summation*, *average* and *weighted average*. As stated in Section 4.2.1, nodes sampled by a weighted root-seeking walk are weighted in the descending order from the PoI node up to the root. The resultant sentence embeddings $u$ and $v$ are concatenated with element-wise difference $|u - v|$ to get the final embedding vector, which is then fed into a softmax classifier for the polarity prediction task.

In order to fine-tune BERT, we make the GraphNLI model end-to-end trainable to update weights during backpropagation such that the produced sentence embeddings are semantically meaningful for the downstream prediction tasks.

## 5 EXPERIMENTS AND RESULTS

In this section, we describe our experimental setup for training the GraphNLI model, the results obtained, and a detailed ablation study to better understand different aspects of the model.

### 5.1 Dataset Preprocessing

As described in Section 3, we use data from online debates conducted on Kialo in order to train and evaluate our model. All discussions crawled from Kialo have a tree structure with a root node that represents the main thesis and each other node is a reply to its parent which either supports or attacks the parent.

As discussed in Section 4.1, we represent Kialo debates as undirected discussion trees. Each edge can have a positive or negative sign, indicating support or attack respectively. As mentioned in Section 3, Kialo debates are typically balanced, with the majority of discussion trees having a fraction of supportive replies between 0.4 and 0.6 with the rest being attacks; this justifies our use of accuracy as the metric to evaluate our polarity classifiers (see Section 5.4). We randomly sample 80% of the Kialo debates into a training set with the remainder serving as a test set. Overall, the training set contains of 259, 499 arguments (replies) in total, while the test set contains 64, 874 arguments in Kialo debates.

### 5.2 Training Details

After preprocessing the Kialo dataset, we use the graph walk techniques described in Section 4.2.1 to capture the neighbourhood and parent contexts for each of the nodes and feed them into our GraphNLI model.

In the case of the weighted root-seeking graph walk, we assign weights to the nodes in the graph walk progressively. We define a discount parameter $\gamma = 0.75$ such that the nodes are weighted in the form of $\gamma^k$, where $k$ is the distance of the node from the given node. So, the largest weight $\gamma^1 = 0.75$ is assigned to the node's parent, and then $\gamma^2$ is assigned to the parent's parent and so on in the graph walk with $L$ nodes. Therefore, the influence of the

ancestor nodes decreases exponentially as we move farther away from the given node towards the root in a graph walk.

We fine-tune GraphNLI model with a softmax classifier objective function and cross-entropy loss for four epochs. We use a batch-size of 16, Adam optimizer with learning rate $2 \times 10^{-5}$, and a linear learning rate warm-up over 10% of the training data.

## 5.3 Baselines

We compare GraphNLI with the following relevant baselines on the classification accuracy.

**Bag-of-Words with Logistic Regression**: A bag-of-words (BoW) model that uses unigram features as input obtained from the arguments/replies in online debates. Then the parent and the child BoW embeddings are concatenated and fed into a Logistic Regression classifier with L2 regularization trained for 100 epochs.

**Prompt embeddings (Rhetorics) with Logistic Regression**: The prompt embedding model [44] infers a vector representation of utterances / arguments in terms of the responses that similar utterances tend to prompt, as well as the rhetorical intentions encapsulated by such utterances / arguments. Once the embedding vectors for the arguments are obtained, the parent and child embeddings are concatenated and input to a Logistic Regression classifier for the final polarity prediction task. The classifier is trained for 100 epochs with L2 regularizaton and cross-entropy loss.

**Sentence-BERT**: S-BERT [38] is a modification of a pre-trained BERT transformer network [19] to derive semantically meaningful sentence embeddings. We use the S-BERT architecture with classification objective function and input the sentence pairs (parent and child arguments) into the model to get their sentence embeddings. Later on, these embeddings are concatenated and fed into a softmax classifier. The S-BERT model is fine-tuned on Kialo training dataset for 4 epochs with a batch-size of 16, the Adam optimizer, and cross-entropy loss function.

**Non-trainable BERT embeddings with graph walks and Multi-layer Perceptron**: For each of the arguments in online debates, their embeddings are derived using a pre-trained BERT model and using CLS-token embeddings. Using different graph walk techniques as described in Section 4.2.1, various neighborhood siblings and parent nodes are sampled for each node, and using their node embeddings, a resulting aggregated embedding is formed using an average aggregation function. These node embeddings are then fed into a multi-layer perceptron (MLP) with two layers and with a softmax objective function for polarity prediction. The initial BERT embeddings are non-trainable. We train the MLP for 50 epochs or until the model converges on Kialo training dataset with batch-size of 16, and the Adam optimizer.

## 5.4 Model Evaluation

We evaluate the performance of GraphNLI model on the test set of Kialo dataset. Since the polarity prediction task is a binary classification problem and that the datasets are roughly balanced between both classes (attacks and supports), we use accuracy as the evaluation metric. We train models with five different random seeds and report their average performances. Table 1 shows the accuracy scores of different models trained on Kialo train set for the polarity prediction task. The baseline model, Bag-of-Words embeddings with Logistic Regression, achieves an accuracy of 67%, while Prompt embeddings (Rhetorics) with Logistic Regression achieves just 61.20% accuracy. The sentence-BERT model trained on Kialo dataset achieves an accuracy of 79.86%. The initial MLP model with non-trainable BERT embeddings and root-seeking graph walk achieves an accuracy of 70.27% which is even worse than the Sentence-BERT.

**Table 1: Accuracy scores of different models trained on Kialo dataset for polarity prediction, discussed in Section 5.4.**

| Model | Accuracy (%) |
|---|---|
| Bag-of-Words + Logistic Regression | 67.00 |
| Prompt Embeddings + Logistic Regression | 61.20 |
| Sentence-BERT with classifier layer | 79.86 |
| BERT Embeddings: Root-seeking Graph Walk + MLP | 70.27 |
| GraphNLI: Root-seeking Graph Walk + Sum | 80.70 |
| GraphNLI: Root-seeking Graph Walk + Avg. | 81.96 |
| GraphNLI: Root-seeking Graph Walk + Weighted Avg. | **82.87** |
| GraphNLI: Biased Root-seeking Random Walk + Sum | 79.95 |
| GraphNLI: Biased Root-seeking Random Walk + Avg. | 80.44 |

Our model, GraphNLI with root-seeking graph walk and averaging node embeddings in the graph walk to get the aggregated node embeddings achieves an overall accuracy of 81.96%, whereas, the model achieves even better accuracy of 82.87% using weighted average node embeddings. Using biased root-seeking random walk and average node embeddings, GraphNLI achieves an accuracy of 80.44%. Clearly, all the variants of GraphNLI model achieve better accuracy scores than all the baselines including sentence-BERT. GraphNLI with root-seeking graph walk and weighted average node embeddings achieves the highest accuracy overall. This shows that global context of the online debates or discussions along with the local context of the argument pairs indeed helps in improving the performance of the model.

The best performing variant, GraphNLI with root-seeking graph walk and weighted average node embeddings, shows that having context of the upstream arguments in the online debates helps the model in predicting argumentative relations of support and attack. Also, weighted average aggregation gives higher weights to the arguments near to the given argument pair in the discussion tree whose polarity needs to be predicted, and progressively reduces the weights when the graph walk moves towards the root.

## 5.5 Ablation Study

We have demonstrated superior performance of the GraphNLI architecture with respect to various baselines in Table 1. In this section, we perform an ablation study and discuss different aspects of the GraphNLI model and intuition behind the choices in order to gain a better understanding of the model.

First, we evaluate different kinds of graph walks by feeding the resultant embeddings obtained with the *average* aggregation strategy into our model and compare their accuracy scores. As shown in Table 1, all the graph walks with GraphNLI architecture perform better than the S-BERT model significantly. S-BERT just considers the argument pairs (node and its parent embeddings), whereas

through graph walks, GraphNLI considers the global context of discussion trees by exploring parents and neighborhoods of a node. Therefore, the global context with the local context of the ongoing discussions indeed helps in predicting the polarities. On comparing different graph walks themselves, we find the Root-seeking Graph Walks performing better than the Biased Root-seeking Random Walks with 1.52% higher accuracy. It shows that upstream text in the discussions (parent and other ancestor nodes) indeed helps the model more than the sibling and child nodes possibly obtained with the biased root-seeking random walk. Intuitively, a person's reply can be influenced by the on-going discussion in the upstream text, but cannot be influenced by future replies (child nodes) or sibling (parallel) replies.

We evaluate different aggregation strategies (*summation*, *average* and *weighted average*) to aggregate the node embeddings of the neighbouring nodes using a Root-seeking Graph Walk. As shown in Table 1, the weighted average aggregation function performs better than the summation and average strategies. This shows that influence of the upstream text (ancestor nodes) decreases as the graph walk moves away from the given node towards the root. Hence, ancestor nodes cannot be weighted equally but instead, progressively in the decreasing order of their distance from the given node.

**Table 2: Accuracy scores of GraphNLI model trained on Kialo dataset with different concatenation techniques using weighted average aggregation, discussed in Section 5.5.**

| Concatenation | Accuracy (%) |
|---|---|
| $(u, v)$ | 76.78 |
| $(u, v, u * v)$ | 82.05 |
| $(u, v, \|u - v\|)$ | **82.87** |
| $(u, v, \|u - v\|, u * v)$ | 82.38 |

We also evaluate different methods for concatenating a node's embedding $u$ with the aggregated embedding of its neighbours $v$ obtained using the root-seeking graph walk. The impact of the concatenation method on the model's performance is significant. As depicted in Table 2, the concatenation of $(u, v, |u-v|)$ works the best. As reported by [38], adding the element-wise multiplication $u * v$ decreased performance. Element-wise absolute difference $|u - v|$ which measures the distance between the two node embeddings is an important component.

Our initial non-end-to-end trainable model as discussed in Section 5.3 in which we keep the node embeddings obtained from the BERT model fixed, performs even worse than the Sentence-BERT. This throws light on the importance of end-to-end training of the model for fine-tuning on specific tasks. After end-to-end training, the model outputs node embeddings that are rich in context suitable for downstream tasks like polarity prediction.

## 6 CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated a novel model, GraphNLI, that quantifies the polarity of an online interaction by building a representation that captures its content and context. GraphNLI derives inspiration from S-BERT, but it is novel in the sense that it captures the context by sampling a discussion tree of a discourse using different strategies based on graph walks (Section 4.2.1). Empirically, we found that a Root Seeking Graph Walk with a weighted average aggregation of the ancestors' contexts is the best strategy in terms of classification accuracy (Table 1). This strategy addresses the shortcomings of previous approaches that only capture the local context features, such as the reply and the post it is replying to. We also showed through an ablation study that information from parent and other ancestor nodes provide more relevant contextual information than siblings and children of the reply node whose polarity is being determined. Furthermore, the importance of ancestor nodes decreases as the distance from the reply node increases.

A framework for polarity prediction, such as GraphNLI, can have a number of applications for Computational Social Science (CSS):
**Understanding conversation health**: Online discussion forums provide a great opportunity for creative and socially positive interactions, such as peer-support for long-term medical problems [26, 35]. On the other hand, many forums have unfortunately become a medium for rampant misinformation [30] and hate [21]. As such, identifying and promoting "healthy" conversations has been identified as an important priority by many (*e.g.* Twitter [25]). Once the polarity of posts is known, it can be used to develop conversation health metrics such as whether a conversation is supporting or acrimonious (*e.g., cf.* [10]).
**Hate speech**: Hate speech on online forums is a common [15] challenge, including in nationally important conversations between citizens and their elected representatives [1]. In other cases, some members of a discussion can be unfairly targeted, for example, misogyny is understood to be an important problem on online forums such as Reddit [23]. Inferring argument polarities at scale can help platforms to detect such problems before they spiral out of control (for example, highly attacking comments towards female participants can be a possible indicator of potential misogyny).
**Detecting filter bubbles**: Democratic conversations on news and social media sites can exhibit partisan tendencies [2, 3, 7, 27]. This can lead to filter bubbles: two (or more) parallel conversations about the same topic, with each conversation consisting of posts largely agreeing with other posts in that conversation, and yet having a large amount of disagreement with the other conversations happening in parallel. Frameworks like GraphNLI could help detect filter bubbles by quantifying agreeability in conversations: *e.g.* if we find that posts reachable from each other also agree with each other (*i.e.,* are supporting), and yet if an imaginary edge is induced between posts in different parts of a conversation (or a different discussion thread), we find that the imaginary edge would be an attack edge, this could be indicative of a filter bubble.
**Understanding multiple viewpoints in online debates**: A novel way to mitigate problems such as filter bubbles in large online debates is to present readers with a balanced sample of all the justified arguments representing the multiple important viewpoints. Once an argumentation framework has been induced from a discussion and the attack/support relation between posts has been established, it is possible to use the tools of argumentation theory to compute the "winning" arguments from all sides of a debate [41–43]: arguments which have not been refuted and are left standing as valid viewpoints (whether one may agree with them or not).

In future work, we hope to apply GraphNLI to some of the above.

Our work partly fixes the gap of inferring polarity from the content and context of an online discourse. However, several questions remain to be answered to make it relevant for wider usage. For instance, we have demonstrated and evaluated our approach on Kialo, a tightly moderated online debating platform. We would, however, like to test its validity on much noisier, weakly moderated discourses, such as those found on Reddit.[8]. Other forums, such as BBC's *Have Your Say*?[9] do not have an explicit threaded reply structure, requiring us to *infer* from the text of a reply which other post it is replying to, prior to applying GraphNLI's graph walk techniques. In this less restrictive user interface, posts may refer to multiple other posts, which in turns means that the reply graph is no longer a tree. However, we note that discussions that are not trees pose no limitations to our graph walk techniques (Section 4.2.1) because there would only be more contexts for the graph walk to sample.

## REFERENCES


[1] Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. 2021. Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '21)*. Association for Computing Machinery, New York, NY, USA, 5–16. https://doi.org/10.1145/3465336.3475113

[2] Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadapoulos, Nishanth Sastry, and Nicolas Kourtellis. 2020. Stop tracking me Bro! Differential Tracking of User Demographics on Hyper-Partisan Websites. In *Proceedings of the The Web Conference (WWW 2020) (WWW '20)*. International World Wide Web Conferences Steering Committee, Taipei, Taiwan, 10 pages.

[3] Vibhor Agarwal, Yash Vekaria, Pushkal Agarwal, Sangeeta Mahapatra, Shounak Set, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. 2021. Under the Spotlight: Web Tracking in Indian Partisan News Websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 26–37.

[4] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[5] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[6] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. 2011. An introduction to argumentation semantics. *The Knowledge Engineering Review* 26, 4 (2011), 365–410.

[7] Shweta Bhatt, Sagar Joglekar, Shehar Bano, and Nishanth Sastry. 2018. Illuminating an Ecosystem of Partisan Websites. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 545–554. https://doi.org/10.1145/3184558.3188725

[8] Johan Bos and Katja Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL RTE challenge*. 26.

[9] Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. *6th International Conference on Computational Models of Argument* 2016 (2016), 21–32.

[10] Gioia Boschi, Anthony P. Young, Sagar Joglekar, Chiara Cammarota, and Nishanth Sastry. 2021. Who Has the Last Word? Understanding How to Sample Online Discussions. *ACM Transactions on the Web (TWEB)* 15, 3 (2021), 1–25.

[11] Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation* 4, 3 (2013), 209–230.

[12] Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis.. In *IJCAI*, Vol. 18. 5427–5433.

[13] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, 378–389.

[14] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2013. Bipolarity in Argumentation Graphs: Towards a Better Understanding. *International Journal of Approximate Reasoning* 54, 7 (2013), 876–899.

[15] Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. Online Hate: Behavioural Dynamics and Relationship with Misinformation. *arXiv preprint arXiv:2105.14005* (2021).

[16] Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. A dataset independent set of baselines for relation prediction in argument mining. *arXiv preprint arXiv:2003.04970* (2020).

[17] Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1374–1379.

[18] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering* 16, 1 (2010), 105–105.

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

[20] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and $n$-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–357.

[21] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing.

[22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 81–90.

[23] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1336–1350.

[24] Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1859–1874.

[25] Twitter Inc. 2022. Healthy conversations. https://about.twitter.com/en/our-priorities/healthy-conversations

[26] Sagar Joglekar, Nishanth Sastry, Neil S Coulson, Stephanie JC Taylor, Anita Patel, Robbie Duschinsky, Amrutha Anand, Matt Jameson Evans, Chris J Griffiths, Aziz Sheikh, et al. 2018. How online communities of people with long-term conditions function and evolve: network analysis of the structure and dynamics of the asthma UK and British lung foundation online communities. *Journal of Medical Internet Research* 20, 7 (2018), e238.

[27] Dmytro Karamshuk, Tetyana Lokot, Oleksandr Pryymak, and Nishanth Sastry. 2016. Identifying Partisan Slant in News Articles and Twitter During Political Crises. In *Social Informatics*, Emma Spiro and Yong-Yeol Ahn (Eds.). Springer International Publishing, Cham, 257–272.

[28] Sebastian Köffer, Dennis M Riehle, Steffen Höhenberger, and Jörg Becker. 2018. Discussing the value of automatic hate speech detection in online debates. *Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany* (2018).

[29] Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*. 42–47.

[30] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. 591–602.

[31] John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics* 45, 4 (2020), 765–818.

[32] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology (TOIT)* 16, 2 (2016), 1–25.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[34] Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials. *Artificial Intelligence in Medicine* (2021), 102098.

[35] Pietro Panzarasa, Christopher J Griffiths, Nishanth Sastry, and Anna De Simoni. 2020. Social medical capital: how patients and caregivers can benefit from online social interactions. *Journal of Medical Internet Research* 22, 7 (2020), e16337.

[36] Iyad Rahwan, Mohammed I Madakkatel, Jean-François Bonnefon, Ruqiyabi N Awan, and Sherief Abdallah. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science* 34, 8 (2010), 1483–1502.

[37] Iyad Rahwan and Guillermo R. Simari. 2009. *Argumentation in Artificial Intelligence*. Vol. 47. Springer.


[8] https://www.reddit.com/, last accessed 22 Jan 2022.

[9] See the comments of, e.g. https://www.bbc.co.uk/news/uk-43253389#comments, last accessed 22 Jan 2022.

[38] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[39] Serena Villata. 2021. Towards assessing natural language argument strength: results and open challenges. http://argstrength2021.argumentationcompetition.org/

[40] Anthony P. Young. 2018. Notes on Abstract Argumentation Theory. *arXiv preprint arXiv:1806.07709* (2018).

[41] Anthony P. Young. 2021. Likes as Argument Strength for Online Debates. In *The Third Workshop on Argument Strength*. Available from http://argstrength2021.argumentationcompetition.org/papers/ArgStrength2021_paper_8.pdf, last accessed 22/1/2022.

[42] Anthony P. Young, Sagar Joglekar, Gioia Boschi, and Nishanth Sastry. 2021. Ranking comment sorting policies in online debates. *Argument & Computation* 12, 2 (2021), 265–285.

[43] Anthony P. Young, Sagar Joglekar, Kiran Garimella, and Nishanth Sastry. 2018. Approximations to truth in online comment networks. In *The Workshop on Argumentation and Society at the 7th International Conference on Computational Models of Argument*. Available from https://nishrs.github.io/publication/young-2018-comma/, last accessed 22/1/2022.

[44] Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking too much? The rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1558–1572.