The Role of Pragmatic and Discourse Context in Determining Argument Impact

Esin Durmus
Cornell University
ed459@cornell.edu

Faisal Ladhak
Amazon
faisall@amazon.com

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Research in the social sciences and psychology has shown that the persuasiveness of an argument depends not only the language employed, but also on attributes of the source/communicator, the audience, and the appropriateness and strength of the argument's claims given the pragmatic and discourse context of the argument. Among these characteristics of persuasive arguments, prior work in NLP does not explicitly investigate the effect of the pragmatic and discourse context when determining argument quality. This paper presents a new dataset to initiate the study of this aspect of argumentation: it consists of a diverse collection of arguments covering 741 controversial topics and comprising over 47,000 claims. We further propose predictive models that incorporate the pragmatic and discourse context of argumentative claims and show that they outperform models that rely only on claim-specific linguistic features for predicting the perceived impact of individual claims within a particular line of argument.

1 Introduction

Previous work in the social sciences and psychology has shown that the impact and persuasive power of an argument depends not only on the language employed, but also on the credibility and character of the communicator (i.e. ethos) (Miller et al., 1976; Chaiken, 1979, 1980); the traits and prior beliefs of the audience (G. Lord et al., 1979; Davies, 1998; Correll et al., 2004; Hullett, 2005); and the pragmatic context in which the argument is presented (i.e. kairos) (Haugtvedt and Wegener, 1994; Joyce and Harwood, 2014).

Research in Natural Language Processing (NLP) has only partially corroborated these findings. One very influential line of work, for example, develops computational methods to automatically determine the linguistic characteristics of persuasive arguments (Habernal and Gurevych, 2016; Tan et al., 2016; Zhang et al., 2016), but it does so without controlling for the audience, the communicator or the pragmatic context.

Very recent work, on the other hand, shows that attributes of both the audience and the communicator constitute important cues for determining argument strength (Lukin et al., 2017; Durmus and Cardie, 2018). They further show that audience and communicator attributes can influence the relative importance of linguistic features for predicting the persuasiveness of an argument. These results confirm previous findings in the social sciences that show a person's perception of an argument can be influenced by his background and personality traits.

To the best of our knowledge, however, no NLP studies explicitly investigate the role of *kairos* — a component of pragmatic context that refers to the context-dependent "timeliness" and "appropriateness" of an argument and its claims within an argumentative discourse — in argument quality prediction. Among the many social science studies of attitude change, the order in which argumentative claims are shared with the audience has been studied extensively: Haugtvedt and Wegener (1994), for example, summarize studies showing that the argument-related claims a person is exposed to beforehand can affect his perception of an alternative argument in complex ways. Joyce and Harwood (2014) similarly find that changes in an argument's context can have a big impact on the audience's perception of the argument.

Some recent studies in NLP have investigated the effect of interactions on the overall persuasive power of posts in social media (Tan et al., 2016; Hidey and McKeown, 2018). However, in social media not all posts have to express arguments or stay on topic (Rakshit et al., 2017), and qualitative evaluation of the posts can be influenced by many

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5668–5678,
Hong Kong, China, November 3–7, 2019.

other factors such as interactions between the individuals (Durmus and Cardie, 2019). Therefore, it is difficult to measure the effect of argumentative pragmatic context alone in argument quality prediction without the effect of these confounding factors using the datasets and models currently available in this line of research.

In this paper, we study the role of kairos on argument quality prediction by examining the individual claims of an argument for their timeliness and appropriateness in the context of a particular line of argument. We define kairos as the sequence of **argumentative** text (e.g. claims) along a particular line of argumentative reasoning.

To start, we present a dataset extracted from *kialo.com* of over 47,000 claims that are part of a diverse collection of arguments on 741 controversial topics. The structure of the website dictates that each argument must present a supporting or opposing claim for its parent claim, and stay within the topic of the main thesis. Rather than being posts on a social media platform, these are community-curated claims. Furthermore, for each presented claim, the audience votes on its impact within the given line of reasoning. Critically then, the dataset includes the argument context for each claim, allowing us to investigate the characteristics associated with impactful arguments.

With the dataset in hand, we propose the task of studying the characteristics of impactful claims by (1) taking the argument context into account, (2) studying the extent to which this context is important, and (3) determining the representation of context that is more effective. To the best of our knowledge, ours is the first dataset that includes claims with both impact votes and the corresponding context of the argument.

2 Related Work

Recent studies in computational argumentation have mainly focused on the tasks of identifying the structure of the arguments such as argument structure parsing (Peldszus and Stede, 2015; Park and Cardie, 2014), and argument component classification (Habernal and Gurevych, 2017; Mochales and Moens, 2011). More recently, there is an increased research interest to develop computational methods that can automatically evaluate qualitative characteristic of arguments, such as their impact and persuasive power (Habernal and Gurevych, 2016; Tan et al., 2016; Kelman, 1961; Burgoon et al., 1975; Chaiken, 1987; Tykocinskl et al., 1994; Dillard and Pfau, 2002; Cialdini, 2007; Durik et al., 2008; Marquart and Naderer, 2016). Consistent with findings in the social sciences and psychology, some of the work in NLP has shown that the impact and persuasive power of the arguments are not simply related to the linguistic characteristics of the language, but also on characteristics the source (ethos) (Durmus and Cardie, 2019) and the audience (Lukin et al., 2017; Durmus and Cardie, 2018). These studies suggest that perception of the arguments can be influenced by the credibility of the source, and the background of the audience.

It has also been shown, in social science studies, that *kairos*, which refers to the "timeliness" and "appropropriateness" of arguments and claims, is important to consider in studies of argument impact and persuasiveness (Haugtvedt and Wegener, 1994; Joyce and Harwood, 2014). One recent study in NLP has investigated the role of argument sequencing in argument persuasion looking at (Hidey and McKeown, 2018) Change My View[1], which is a social media platform where users post their views, and challenge other users to present arguments in an attempt to change their them. However, as stated in (Rakshit et al., 2017) many posts on social media platforms either do not express an argument, or diverge from the main topic of conversation. Therefore, it is difficult to measure the effect of pragmatic context in argument impact and persuasion, without confounding factors from using noisy social media data. In contrast, we provide a dataset of claims along with their structured argument path, which only consists of claims and corresponds to a particular line of reasoning for the given controversial topic. This structure enables us to study the characteristics of impactful claims, accounting for the effect of the pragmatic context.

Consistent with previous findings in the social sciences, we find that incorporating pragmatic and discourse context is important in computational studies of persuasion, as predictive models that with the context representation outperform models that only incorporate claim-specific linguistic features, in predicting the impact of a claim. Such a system that can predict the impact of a claim given an argumentative discourse, for example, could potentially be employed by argument retrieval and

[1]https://www.reddit.com/r/changemyview/.

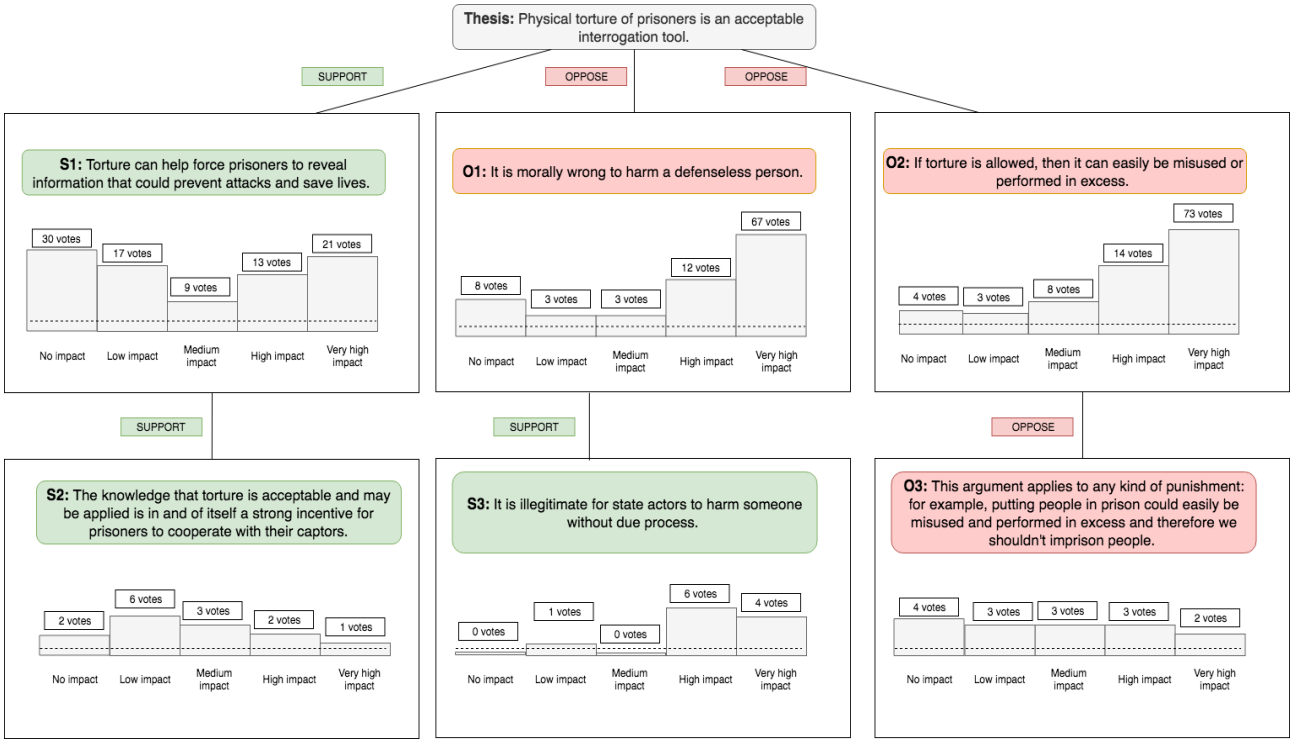

Figure 1: Example partial argument tree with claims and corresponding impact votes for the thesis "PHYSICAL TORTURE OF PRISONERS IS AN ACCEPTABLE INTERROGATION TOOL.".

generation models which aims to pick or generate the most appropriate possible claim given the discourse.

3 Dataset

Claims and impact votes. We collected 47,219 claims from *kialo.com*[2][3] for 741 controversial topics and their corresponding impact votes. Impact votes are provided by the users of the platform to evaluate how impactful a particular claim is. Users can pick one of 5 possible impact labels for a particular claim: NO IMPACT, LOW IMPACT, MEDIUM IMPACT, HIGH IMPACT and VERY HIGH IMPACT. While evaluating the impact of a claim, users have access to the full argument context and therefore, they can assess how impactful a claim is in the given context of an argument. An interesting observation is that, in this dataset, the same claim can have different impact labels depending on the context in which it is presented.

Figure 1 shows a partial **argument tree** for the argument **thesis** "PHYSICAL TORTURE OF PRISONERS IS AN ACCEPTABLE INTERROGATION TOOL.". Each node in the argument tree corresponds to a claim, and these argument trees are constructed and edited collaboratively by the users of the platform.

Except the thesis, every claim in the argument tree either opposes or supports its parent claim. Each path from the root to leaf nodes corresponds to an **argument path** which represents a particular line of reasoning on the given controversial topic.

Moreover, each claim has **impact votes** assigned by the users of the platform. The impact votes evaluate how impactful a claim is within its context, which consists of its predecessor claims from the thesis of the tree. For example, claim **O1** "IT IS MORALLY WRONG TO HARM A DEFENSELESS PERSON" is an opposing claim for the thesis and it is an IMPACTFUL CLAIM since most of its impact votes belong to the category of VERY HIGH IMPACT. However, claim **S3** "IT IS ILLEGITIMATE FOR STATE ACTORS TO HARM SOMEONE WITHOUT THE PROCESS" is a supporting claim for its parent **O1** and it is a less impactful claim since most of the impact votes belong to the NO IMPACT and LOW IMPACT categories.

Distribution of impact votes. The distribution of claims with the given range of number of im-

[2]The data is collected from this website in accordance with the terms and conditions.

[3]There is prior work by Durmus et al. (2019) which created a dataset of argument trees from *kialo.com*. That dataset, however, does not include any impact labels.

# impact votes	# claims
$[3, 5)$	4,495
$[5, 10)$	5,405
$[10, 15)$	5,338
$[15, 20)$	2,093
$[20, 25)$	934
$[25, 50)$	992
$[50, 333)$	255

Table 1: Number of claims for the given range of number of votes. There are 19,512 claims in the dataset with 3 or more votes. Out of the claims with 3 or more votes, majority of them have 5 or more votes.

pact votes are shown in Table 1. There are 19,512 claims in total with 3 or more votes. Out of the claims with 3 or more votes, majority of them have 5 or more votes. We limit our study to the claims with at least 5 votes to have a more reliable assignment for the accumulated impact label for each claim.

Impact label statistics. Table 3 shows the distribution of the number of votes for each of the impact categories. The claims have $241,884$ total votes. The majority of the impact votes belong to MEDIUM IMPACT category. We observe that users assign more HIGH IMPACT and VERY HIGH IMPACT votes than LOW IMPACT and NO IMPACT votes respectively. When we restrict the claims to the ones with at least 5 impact votes, we have $213,277$ votes in total[4].

Agreement for the impact votes. To determine the agreement in assigning the impact label for a particular claim, for each claim, we compute the percentage of the votes that are the same as the majority impact vote for that claim. Let c_i denote the count of the claims with the class labels C=[NO IMPACT, LOW IMPACT, MEDIUM IMPACT, HIGH IMPACT, VERY HIGH IMPACT] for the impact label l at index i.

$$\text{Agreement} = 100 * \frac{\max_{0 \leq i \leq 4} c_i}{\sum_{i=0}^{4} c_i}\% \quad (1)$$

For example, for claim S1 in Figure 1, the agreement score is $100 * \frac{30}{90}\% = 33.33\%$ since the majority class (NO IMPACT) has 30 votes and there are 90 impact votes in total for this particular claim. We compute the agreement score for the cases where (1) we treat each impact label separately (5-class case) and (2) we combine the classes HIGH IMPACT and VERY HIGH IMPACT into a one class: IMPACTFUL and NO IMPACT and LOW IMPACT into a one class: NOT IMPACTFUL (3-class case).

Table 2 shows the number of claims with the given agreement score thresholds when we include the claims with at least 5 votes. We see that when we combine the low impact and high impact classes, there are more claims with high agreement score. This may imply that distinguishing between no impact-low impact and high impact-very high impact classes is difficult. To decrease the sparsity issue, in our experiments, we use 3-class representation for the impact labels. Moreover, to have a more reliable assignment of impact labels, we consider only the claims with have more than 60% agreement.

Context. In an argument tree, the claims from the thesis node (root) to each leaf node, form an argument path. This argument path represents a particular line of reasoning for the given thesis. Similarly, for each claim, all the claims along the path from the thesis to the claim, represent the **context** for the claim. For example, in Figure 1, the context for **O1** consists of only the thesis, whereas the context for **S3** consists of both the thesis and **O1** since **S3** is provided to support the claim **O1** which is an opposing claim for the thesis.

The claims are not constructed independently from their context since they are written in consideration with the line of reasoning so far. In most cases, each claim elaborates on the point made by its parent and presents cases to support or oppose the parent claim's points. Similarly, when users evaluate the impact of a claim, they consider if the claim is timely and appropriate given its context. There are cases in the dataset where the same claim has different impact labels, when presented within a different context. Therefore, we claim that it is not sufficient to only study the linguistic characteristic of a claim to determine its impact, but it is also necessary to consider its context in determining the impact.

Context length (C_l) for a particular claim *C* is defined by number of claims included in the argument path starting from the thesis until the claim *C*. For example, in Figure 1, the context length for **O1** and **S3** are 1 and 2 respectively. Table 4 shows number of claims with the given range of con-

[4]26,998 of them NO IMPACT, 33,789 of them LOW IMPACT, 55,616 of them MEDIUM IMPACT, 47,494 of them HIGH IMPACT and 49,380 of them VERY HIGH IMPACT.

	3-class case	5-class case
Agreement score	Number of claims	Number of claims
$> 50\%$	10,848	7,304
$> 60\%$	7,386	4,329
$> 70\%$	4,412	2,195
$> 80\%$	2,068	840

Table 2: Number of claims, with at least 5 votes, above the given threshold of agreement percentage for 3-class and 5-class cases. When we combine the low impact and high impact classes, there are more claims with high agreement score.

Impact label	# votes- all claims
No impact	32,681
Low impact	37,457
Medium impact	60,136
High impact	52,764
Very high impact	58,846
Total # votes	241,884

Table 3: Number of votes for the given impact label. There are $241,884$ total votes and majority of them belongs to the category MEDIUM IMPACT.

Context length	# claims
1	1,524
2	1,977
3	1,181
$[4, 5]$	1,436
$(5, 10]$	1,115
> 10	153

Table 4: Number of claims for the given range of context length, for claims with more than 5 votes and an agreement score greater than 60%.

text length for the claims with more than 5 votes and 60% agreement score. We observe that more than half of these claims have 3 or higher context length.

4 Methodology

4.1 Hypothesis and Task Description

Similar to prior work, our aim is to understand the characteristics of impactful claims in argumentation. However, we **hypothesize** that the qualitative characteristics of arguments is not independent of the context in which they are presented. To understand the relationship between argument context and the impact of a claim, we aim to incorporate the context along with the claim itself in our predictive models.

Prediction task. Given a claim, we want to predict the impact label that is assigned to it by the users: NOT IMPACTFUL, MEDIUM IMPACT, or IMPACTFUL.

Preprocessing. We restrict our study to claims with at least 5 or more votes and greater than 60% agreement, to have a reliable impact label assignment. We have $7,386$ claims in the dataset satisfying these constraints[5]. We see that the impact class IMPACFUL is the majority class since around 58% of the claims belong to this category.

For our experiments, we split our data to train (70%), validation (15%) and test (15%) sets.

4.2 Baseline Models

4.2.1 Majority

The majority baseline assigns the most common label of the training examples (HIGH IMPACT) to every test example.

4.2.2 SVM with RBF kernel

Similar to (Habernal and Gurevych, 2016), we experiment with SVM with RBF kernel, with features that represent (1) the simple characteristics of the argument tree and (2) the linguistic characteristics of the claim.

The features that represent the simple characteristics of the claim's argument tree include the distance and similarity of the claim to the thesis, the similarity of a claim with its parent, and the impact votes of the claim's parent claim. We encode the similarity of a claim to its parent and the thesis claim with the cosine similarity of their tf-idf vectors. The distance and similarity metrics aim to model whether claims which are more similar (i.e. potentially more topically relevant) to their parent claim or the thesis claim, are more impactful.

[5]We have 1,633 NOT IMPACTFUL, 1,445 MEDIUM IMPACT and 4,308 IMPACFUL claims.

We encode the quality of the parent claim as the number of votes for each impact class, and incorporate it as a feature to understand if it is more likely for a claim to impactful given an impactful parent claim.

Linguistic features. To represent each claim, we extracted the linguistic features proposed by (Habernal and Gurevych, 2016) such as tf-idf scores for unigrams and bigrams, ratio of quotation marks, exclamation marks, modal verbs, stop words, type-token ratio, hedging (Hyland, 1998), named entity types, POS n-grams, sentiment (Hutto and Gilbert, 2014) and subjectivity scores (Wilson et al., 2005), spell-checking, readibility features such as *Coleman-Liau* (Coleman and Liau, 1975), *Flesch* (Flesch, 1948), argument lexicon features (Somasundaran et al., 2007) and surface features such as word lengths, sentence lengths, word types, and number of complex words[6].

4.2.3 FastText

Joulin et al. (2017) introduced a simple, yet effective baseline for text classification, which they show to be competitive with deep learning classifiers in terms of accuracy. Their method represents a sequence of text as a bag of n-grams, and each n-gram is passed through a look-up table to get its dense vector representation. The overall sequence representation is simply an average over the dense representations of the bag of n-grams, and is fed into a linear classifier to predict the label. We use the code released by Joulin et al. (2017) to train a classifier for argument impact prediction, based on the claim text[7].

4.2.4 BiLSTM with Attention

Another effective baseline (Zhou et al., 2016; Yang et al., 2016) for text classification consists of encoding the text sequence using a bidirectional Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), to get the token representations in context, and then attending (Luong et al., 2015) over the tokens to get the sequence representation. For the query vector for attention, we use a learned context vector, similar to Yang et al. (2016). We picked our hyperparameters based on performance on the validation set, and report our results for the best set of hyperparameters[8]. We initialized our word embeddings with glove vectors (Pennington et al., 2014) pre-trained on Wikipedia + Gigaword, and used the Adam optimizer (Kingma and Ba, 2015) with its default settings.

4.3 Fine-tuned BERT model

Devlin et al. (2018) fine-tuned a pre-trained deep bi-directional transformer language model (which they call BERT), by adding a simple classification layer on top, and achieved state of the art results across a variety of NLP tasks. We employ their pre-trained language models for our task and compare it to our baseline models. For all the architectures described below, we finetune for 10 epochs, with a learning rate of 2e-5. We employ an early stopping procedure based on the model performance on a validation set.

4.3.1 Claim with no context

In this setting, we attempt to classify the impact of the claim, based on the text of the claim only. We follow the fine-tuning procedure for sequence classification detailed in (Devlin et al., 2018), and input the claim text as a sequence of tokens preceded by the special [CLS] token and followed by the special [SEP] token. We add a classification layer on top of the BERT encoder, to which we pass the representation of the [CLS] token, and fine-tune this for argument impact prediction.

4.3.2 Claim with parent representation

In this setting, we use the parent claim's text, in addition to the target claim text, in order to classify the impact of the target claim. We treat this as a sequence pair classification task, and combine both the target claim and parent claim as a single sequence of tokens, separated by the special separator [SEP]. We then follow the same procedure above, for fine-tuning.

4.3.3 Incorporating larger context

In this setting, we consider incorporating a larger context from the discourse, in order to assess the impact of a claim. In particular, we consider up

[6] We pick the parameters for SVM model according to the performance validation split, and report the results on the test split.

[7] We used maxNgram legnth of 2, learning rate of 0.8, num epochs of 15, vector dim of 300. We also used the pre-trained 300-dim wiki-news vectors made available on the fastText website.

[8] Our final hyperparams were: 100-dim word embedding, 100-dim context vector, 1 layer BiLSTM with 64 units, trained for 40 epochs with early stopping based on validation performance.

	Precision	Recall	F1
Majority	19.43	33.33	24.55
SVM with RBF Kernel			
Distance from the thesis	27.42	33.53	26.05
Parent quality	58.11	47.85	46.61
Linguistic features	**65.67**	38.58	35.42
BiLSTM with Attention	$46.50_{\pm 0.28}$	$46.35_{\pm 0.99}$	$46.22_{\pm 0.58}$
FastText	$51.18_{\pm 0.80}$	$46.09_{\pm 0.64}$	$47.06_{\pm 0.70}$
BERT models			
Claim only	$53.24_{\pm 1.07}$	$50.93_{\pm 2.01}$	$51.53_{\pm 1.53}$
Claim + Parent	$55.79_{\pm 1.72}$	$53.54_{\pm 2.09}$	$54.00_{\pm 1.79}$
Claim + Context$_f(2)$	$56.57_{\pm 0.85}$	$54.76_{\pm 1.71}$	$55.18_{\pm 0.99}$
Claim + Context$_f(3)$	$57.19_{\pm 0.92}$	$\mathbf{55.77_{\pm 1.05}}$	$\mathbf{55.98_{\pm 0.70}}$
Claim + Context$_f(4)$	$57.09_{\pm 1.71}$	$55.31_{\pm 1.09}$	$55.72_{\pm 1.14}$
Claim + Context$_{gru}(4)$	$54.95_{\pm 2.00}$	$51.55_{\pm 1.27}$	$52.37_{\pm 1.26}$
Claim + Context$_a(4)$	$56.60_{\pm 0.52}$	$54.55_{\pm 0.57}$	$54.65_{\pm 0.33}$

Table 5: Results for the baselines and the BERT models with and without the context. Best performing model is BERT with the representation of previous 3 claims in the path along with the claim representation itself. We run the models 5 times and we report the mean and standard deviation.

to four previous claims in the discourse (for a total context length of 5). We attempt to incorporate larger context into the BERT model in three different ways.

Flat representation of the path. The first, simple approach is to represent the entire path (claim + context) as a single sequence, where each of the claims is separated by the [SEP] token. BERT was trained on sequence pairs, and therefore the pre-trained encoders only have two segment embeddings (Devlin et al., 2018). So to fit multiple sequences into this framework, we indicate all tokens of the target claim as belonging to segment A and the tokens for all the claims in the discourse context as belonging to segment B. This way of representing the input, requires no additional changes to the architecture or retraining, and we can just finetune in a similar manner as above. We refer to this representation of the context as a flat representation, and denote the model as Context$_f(i)$, where i indicates the length of the context that is incorporated into the model.

Attention over context. Recent work in incorporating argument sequence in predicting persuasiveness (Hidey and McKeown, 2018) has shown that hierarchical representations are effective in representing context. Similarly, we consider hierarchical representations for representing the discourse. We first encode each claim using the pre-trained BERT model as the claim encoder, and use the representation of the [CLS] token as claim representation. We then employ dot-product attention (Luong et al., 2015), to get a weighted representation for the context. We use a learned context vector as the query, for computing attention scores, similar to Yang et al. (2016). The attention score α_c is computed as shown below:

$$\alpha_c = \frac{exp(V_c^T V_l)}{\sum_{c \in D} exp(V_c^T V_l)} \quad (2)$$

Where V_c is the claim representation that was computed with the BERT encoder as described above, V_l is the learned context vector that is used for computing attention scores, and D is the set of claims in the discourse. After computing the attention scores, the final context representation v_d is computed as follows:

$$V_d = \sum_{c \in D} \alpha_c V_c \quad (3)$$

We then concatenate the context representation with the target claim representation $[V_d, V_r]$ and pass it to the classification layer to predict the quality. We denote this model as Context$_a(i)$.

GRU to encode context Similar to the approach above, we consider a hierarchical representation for representing the context. We compute the claim representations, as detailed above, and we then feed the discourse claims' representations (in sequence) into a bidirectional Gated Recurrent

	$C_l = 1$	$C_l = 2$	$C_l = 3$	$C_l = 4$
BERT models				
Claim only	$48.61_{\pm 3.16}$	$53.15_{\pm 1.95}$	$54.51_{\pm 1.91}$	$50.89_{\pm 2.95}$
Claim + Parent	$51.49_{\pm 2.63}$	$54.78_{\pm 2.95}$	$54.94_{\pm 2.72}$	$51.94_{\pm 2.59}$
Claim + Context$_f(2)$	$52.84_{\pm 2.55}$	$53.77_{\pm 1.00}$	$55.24_{\pm 2.52}$	$57.04_{\pm 1.19}$
Claim + Context$_f(3)$	$\mathbf{54.88_{\pm 2.49}}$	$54.71_{\pm 1.74}$	$52.93_{\pm 2.07}$	$\mathbf{58.17_{\pm 1.89}}$
Claim + Context$_f(4)$	$54.47_{\pm 2.95}$	$\mathbf{54.88_{\pm 1.53}}$	$\mathbf{57.11_{\pm 3.38}}$	$57.02_{\pm 2.22}$

Table 6: F1 scores of each model for the claims with various context length values.

Unit (GRU) (Cho et al., 2014), to compute the context representation. We concatenate this with the target claim representation and use this to predict the claim impact. We denote this model as Context$_{gru}(i)$.

5 Results and Analysis

Table 5 shows the macro precision, recall and F1 scores for the baselines as well as the BERT models with and without context representations[9].

We see that *parent quality* is a simple yet effective feature and SVM model with this feature can achieve significantly higher ($p < 0.001$)[10] F1 score (46.61%) than *distance from the thesis* and *linguistic features*. Claims with higher impact parents are more likely to be have higher impact. *Similarity with the parent and thesis* is not significantly better than the *majority* baseline. Although the BiLSTM model with attention and FastText baselines performs better than the SVM with *distance from the thesis* and *linguistic features*, it has similar performance to the *parent quality* baseline.

We find that the BERT model with *claim only* representation performs significantly better ($p < 0.001$) than the baseline models. Incorporating the *parent representation* only along with the *claim representation* does not give significant improvement over representing the claim only. However, *incorporating the flat representation of the larger context* along with the claim representation consistently achieves significantly better ($p < 0.001$) performance than the claim representation alone. Similarly, *attention representation* over the context with the learned query vector achieves significantly better performance then the *claim representation* only ($p < 0.05$).

We find that the *flat representation* of the context achieves the highest F1 score. It may be more difficult for the models with a larger number of parameters to perform better than the *flat representation* since the dataset is small. We also observe that modeling 3 claims on the argument path before the target claim achieves the best F1 score (55.98%).

To understand for what kinds of claims the best performing contextual model is more effective, we evaluate the BERT model with *flat context representation* for claims with context length values 1, 2, 3 and 4 separately. Table 6 shows the F1 score of the BERT model without context and with *flat context representation* with different lengths of context. For the claims with context length 1, adding Context$_f(3)$ and Context$_f(4)$ representation along with the claim achieves significantly better ($p < 0.05$) F1 score than modeling the *claim only*. Similarly for the claims with context length 3 and 4, Context$_f(4)$ and Context$_f(3)$ perform significantly better than BERT with *claim only* (($p < 0.05$) and ($p < 0.01$) respectively). We see that models with larger context are helpful even for claims which have limited context (e.g. $C_l = 1$). This may suggest that when we train the models with larger context, they learn how to represent the claims and their context better.

6 Conclusion

In this paper, we present a dataset of claims with their corresponding impact votes, and investigate the role of argumentative discourse context in argument impact classification. We experiment with various models to represent the claims and their context and find that incorporating the context information gives significant improvement in predicting argument impact. In our study, we find that *flat representation* of the context gives the best improvement in the performance and our analysis indicates that the contextual models perform better even for the claims with limited context.

[9]For the models that result in different scores with different random seed, we run them 5 times and report the mean and standard deviation.

[10]We perform two-sided t test for significance analysis.

7 Acknowledgements

This work was supported in part by NSF grants IIS-1815455 and SES-1741441. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

Michael Burgoon, Stephen B Jones, and Diane Stewart. 1975. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity1. *Human Communication Research*, 1(3):240–256.

S. Chaiken. 1979. Communicator physical attractiveness and persuasion.

Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39:752–766.

Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39. Hillsdale, NJ: Lawrence Erlbaum.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Robert B. Cialdini. 2007. Influence: The psychology of persuasion.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283 – 284.

Joshua Correll, Steven J. Spencer, and Mark P. Zanna. 2004. An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength.

MF Davies. 1998. Dogmatism and belief formation: Output interference in the processing of supporting and contradictory cognitions. *Journal of Personality and Social Psychology*, 75:456–466.

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*.

James Price Dillard and Michael Pfau. 2002. *The persuasion handbook: Developments in theory and practice*. Sage Publications.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3):217–234.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. 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 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, WWW '19, pages 2701–2707, New York, NY, USA. ACM.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221 – 233.

Charles G. Lord, Lee Ross, and Mark Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *American Journal of Computational Linguistics*, 43(1):125–179.

Curtis P. Haugtvedt and Duane T. Wegener. 1994. Message Order Effects in Persuasion: An Attitude Strength Perspective. *Journal of Consumer Research*, 21(1):205–218.

Christopher Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Craig R. Hullett. 2005. The impact of mood on persuasion: A meta-analysis. *Communication Research*, 32(4):423–442.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nick Joyce and Jake Harwood. 2014. Context and identification in persuasive mass communication. *Journal of Media Psychology: Theories, Methods, and Applications*, 26:50.

Herbert C Kelman. 1961. Processes of opinion change. *Public opinion quarterly*, 25(1):57–78.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Franziska Marquart and Brigitte Naderer. 2016. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, pages 231–242. Springer Fachmedien Wiesbaden, Wiesbaden.

Norman Miller, Geoffrey Maruyama, Rex J. Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *Journal of Personality and Social Psychology*, 34(4):615–624.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. *CoRR*, abs/1709.03167.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Orit Tykocinskl, E Tory Higgins, and Shelly Chaiken. 1994. Message framing, self-discrepancies, and yielding to persuasive messages: The motivational significance of psychological situations. *Personality and Social Psychology Bulletin*, 20(1):107–115.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 136–141, San Diego, California. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.