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 on 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
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; Tykocinski 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 “appropriateness” 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\textsuperscript{2} 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 O\textsubscript{1} “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 S\textsubscript{3} “IT IS ILLEGIMATE FOR STATE ACTORS TO HARM SOMEONE WITHOUT THE PROCESS” is a supporting claim for its parent O\textsubscript{1} 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-
Impact 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 \(C_i = \text{[NO IMPACT, LOW IMPACT, MEDIUM IMPACT, HIGH IMPACT, VERY HIGH IMPACT]}\) for the impact label \(l\) at index \(i\).

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

For example, for claim S1 in Figure 1, the agreement score is 100 \(	imes\) \(\frac{30}{90}\) \(=\) 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 \((C_i)\) 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-

<table>
<thead>
<tr>
<th># impact votes</th>
<th># claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3, 5)</td>
<td>4,495</td>
</tr>
<tr>
<td>[5, 10)</td>
<td>5,405</td>
</tr>
<tr>
<td>[10, 15)</td>
<td>5,338</td>
</tr>
<tr>
<td>[15, 20)</td>
<td>2,093</td>
</tr>
<tr>
<td>[20, 25)</td>
<td>934</td>
</tr>
<tr>
<td>[25, 50)</td>
<td>992</td>
</tr>
<tr>
<td>[50, 333)</td>
<td>255</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Impact label</th>
<th># votes- all claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>No impact</td>
<td>32,681</td>
</tr>
<tr>
<td>Low impact</td>
<td>37,457</td>
</tr>
<tr>
<td>Medium impact</td>
<td>60,136</td>
</tr>
<tr>
<td>High impact</td>
<td>52,764</td>
</tr>
<tr>
<td>Very high impact</td>
<td>58,846</td>
</tr>
<tr>
<td>Total # votes</td>
<td>241,884</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Context length</th>
<th># claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,524</td>
</tr>
<tr>
<td>2</td>
<td>1,977</td>
</tr>
<tr>
<td>3</td>
<td>1,181</td>
</tr>
<tr>
<td>[4, 5)</td>
<td>1,436</td>
</tr>
<tr>
<td>(5, 10]</td>
<td>1,115</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>153</td>
</tr>
</tbody>
</table>

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.

### 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.

---

We have 1,633 NOT IMPACTFUL, 1,445 MEDIUM IMPACT and 4,308 IMPACTFUL 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, readability 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.

### 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.

### 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. We initialized our word embeddings with glove vectors (Pennington et al., 2014) pretrained 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 fine-tune 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 length 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.
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 $\alpha_c$ is computed as shown below:

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

Where $V_c$ is the claim representation that was computed with the BERT encoder as described above, $V_i$ 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:

$$v_d = \sum_{c \in D} \alpha_c v_c$$

We then concatenate the context representation with the target claim representation $[V_d, V_c]$ 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

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

Table 5: Results for the baselines and the BERT models with and without the context. Best performing model is the 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 $\alpha_c$ is computed as shown above:

We then concatenate the context representation with the target claim representation $[V_d, V_c]$ 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

5678
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 \text{Context}_{gru}(t).

### 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 \textit{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 \textit{distance from the thesis} and \textit{linguistic features}. Claims with higher impact parents are more likely to be have higher impact. \textit{Similarity with the parent and thesis} is not significantly better than the \textit{majority} baseline. Although the BiLSTM model with attention and FastText baselines performs better than the SVM with \textit{distance from the thesis} and \textit{linguistic features}, it has similar performance to the \textit{parent quality} baseline.

We find that the BERT model with \textit{claim only} representation performs significantly better \((p < 0.001)\) than the baseline models. Incorporating the \textit{parent representation} only along with the \textit{claim representation} does not give significant improvement over representing the claim only. However, incorporating the \textit{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, \textit{attention representation} over the context with the learned query vector achieves significantly better performance then the \textit{claim representation} only \((p < 0.05)\).

We find that the \textit{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 \textit{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 \textit{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 \textit{flat context representation} with different lengths of context. For the claims with context length 1, adding \textit{Context}_f(3) and \textit{Context}_f(4) representation along with the claim achieves significantly better \((p < 0.05)\) F1 score than modeling the \textit{claim only}. Similarly for the claims with context length 3 and 4, \textit{Context}_f(4) and \textit{Context}_f(3) perform significantly better than BERT with \textit{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 \textit{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.
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


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


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


