ArgAnalysis35K: A Large-Scale Dataset for Argument Quality Analysis

Omkar Joshi, Priya Pitre, Dr. Y.V. Haribhakta

College of Engineering, Pune
omkarjj98@gmail.com,priya.pitre@gmail.com, ybl.comp@coep.ac.in

Abstract

Argument Quality Detection is an emerging field in NLP which has seen significant recent development. However, existing datasets in this field suffer from a lack of quality, quantity and diversity of topics and arguments, specifically the presence of vague arguments that are not persuasive in nature. In this paper, we leverage a combined experience of 10+ years of Parliamentary Debating to create a dataset that covers significantly more topics and has a wide range of sources to capture more diversity of opinion. With 34,890 high-quality arguments, this is also the largest dataset of its kind to our knowledge. In addition to this contribution, we introduce an innovative argument scoring system based on instance-level annotator reliability and propose a quantitative model of scoring the relevance of arguments to a range of topics.

Introduction

Parliamentary Debate is an extemporaneous form of debating. One of the major intersections of Natural Language Processing and Debating was IBM Project Debater (Slonim et al. 2021), an end-to-end system that mines arguments in a text (Ein-Dor et al. 2019; Toledo-Ronen et al. 2018), determines argument quality (Toledo et al. 2019), and through a combination of modules can debate against a human being. The purpose of this paper is to propose a new dataset that adds a new dimension to the field of argument quality detection in the context of parliamentary debating, eventually enabling the creation of a system that can beat a human debater in a Parliamentary debate.

The dimension that we introduce here is a detailed explanation of why the argument made is true, applicable or impactful, henceforth referred to as "analysis". Analysis is defined as logical links provided to defend a statement, an example of which can be seen in table 2. This can be compared against just arguments, as implemented by (Slonim et al. 2021) seen in table 1. Analysis is important in Parliamentary Debating as a way to defend and analyse arguments as explained in Bazari et al. (2015). The concept of analysis as logically linked statements therefore is an important improvement to the claim-premise concept that is specifically applicable to Parliamentary Debating and that is what we wish to formalize through this paper.

Argument	Motions
a child is still growing, physically and mentally, cosmetic surgery should not be considered until they are an adult and able to make these decisions	We should ban cosmetic surgery for minors
Racial profiling unfairly targets minorities and poor	We should end racial profiling

Table 1: Arguments with score 1 (highest scored arguments) from IBM-30K, a dataset that just collects arguments

Argument relevance is an important indicator of persuasiveness according to Paglieri and Castelfranchi (2014). In a parliamentary debating context, the same argument can be applied to a variety of topics and can be differently persuasive for each topic. Arguments like "accountability is important" can be used in debates about governments, churches, corporations, schools, etc. Similarly, arguments that deal with the premise of free speech being important can be used to defend free speech for members of the LGBTQ community, as well as to defend people's right to protest against a corporation. The quantification of relevance of the argument to the topic under discussion is defined as the relevance model which attempts to capture this complexity.

Application of Instance-based annotator reliability to argumentation is another important contribution described in this paper. Some annotators might know a lot more about art than about the criminal justice system, hence might weigh certain arguments as more or less persuasive using their knowledge; secondly, because of the element of bias that comes in when ranking arguments. Annotators might be biased about a certain argument on race, for example, because of the strong sentiments they feel towards them in their daily life, but they may not be biased when judging an argument

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

on art. We propose a system that enables us to keep the scores of these annotators instead of dropping them, like previous systems have, and show how this leads to a better overall dataset with a more uniform distribution of scores. The dataset is crucial to designing systems that can interact efficiently with humans. Arguments generated with this system can analyze arguments better, and create effective rebuttals using high scoring arguments of the other side. The dataset can also be used to judge a debate by assigning scores to arguments as per their level. Any interactive system, such as IBMs Project Debater needs this dataset as a preliminary base to analyze and win debates with a human.

In summary, our major contributions detailed in this paper are: (1) Argument-analysis pairs collected from a variety of sources on a variety of topics; (2) Introduction of a relevance model that enables the use of multiple arguments in different contexts; (3) Introduction of an instance based annotator scoring system that reduces bias and makes argument scores more accurate.

Argument	Analysis
African American groups should fight for economic reparations from the government.	Reparations are required because African Americans were asked to pay equal taxes while being treated unequally with laws such as Jim Crow laws, $\frac{3}{5}$ citizen rule, etc.
Racial appearance changes should be banned because it leads to discrimination.	Anti discrimination legislation is prefaced on the fact that all races should be treated equally because race is something you cannot change, this is undermined when the government allows changing of race.

Table 2: ArgAnalysis35K argument-analysis pairs with score 1 (highest scored arguments), showing a dataset with argument and analysis
sa

Related Works

Empirical methods rely on creating a particular style of a dataset to assign a score to an argument. There have been several datasets in the field of argument quality using empirical methods that focus on finding arguments and evidence. Roush and Balaji (2020) collects policy arguments and evidence from National Speech and Debate association, while Hua and Wang (2017) categorises arguments into different types like study, factual, opinion, and finds supporting statements for the same. Our work differs from these in several ways: first, the type of evidence used in these papers are either expert citations ("Dr. x recommends y"), results of studies ("According to the 2016 study.."), or opinions of laymen ("In my childhood.."). These are all different from the analysis that we propose, which follows a logical path to reach a conclusion, as seen in Parliamentary Debates ("Cryptocurrency is volatile because companies don't hold it with the intention to make long term profit, which results in no stabilising force being created in the market"). Secondly, these studies aim to find supporting statements, however no quantitative scoring metric has been assigned to the supporting analysis, a problem we solve by giving quantitative scores to both arguments and analysis. Other methods like the one proposed by Persing and Ng (2017) and Habernal and Gurevych (2016a) learn the reasons that an argument is persuasive or non persuasive to improve upon them, all of which provides theoretical reasoning but no quantitative score.

Toledo et al. (2019) and Gretz et al. (2020a) have created IBMRank and IBMRank30K, which contains arguments labelled for quality. Our work is different from theirs in several ways: first, we provide analysis points to arguments which helps us get higher quality arguments from annotators as they are asked to defend their argument without just stating it, and it gives insight into why an argument is persuasive (whether it is persuasive by itself or if the following analysis makes it persuasive) by providing two separate scores. Secondly, these datasets are composed of arguments for random topics that do not cover the diversity of the topics encountered in debating, which is a problem we aim to solve by using 100+ topics covering every genre as stated in multiple sources. Lastly, this dataset is larger in volume than both of these works, consisting of 35K argument-analysis pairs. The methods used to collect data vary for several datasets, some using policy debate arguments from the NSDA (Roush and Balaji 2020), crowdsourcing (IBMs Speech by Crowd), Reddit (Tan et al. 2016). The common factor with all these methods is that they rely on arguments generated either by non-debaters or by crowdsourcing it entirely without knowing the quality of annotators, hence creating a lack of high-quality arguments and variety of arguments.

Lastly, a major contribution in this work is the proposal of a relevance model. Wachsmuth et al. (2017) suggested a model that decomposes quality to 15 dimensions to determine the qualities that make an argument persuasive. They discover that relevance is an important factor that determines argument quality. Gretz et al. (2020a) uses this as the basis to discover that Global Relevance (how related an argument is to the topic) has the highest difference between low and high scoring arguments, hence proving that it is the most important factor that determined how persuasive annotators found it. We use this theory as the basis to create a relevance model that judges this quantitatively. Wachsmuth, Stein, and Ajjour (2017) finds relevance using the number of other arguments that use it as a premise. Our method is different from this as it does not depend on other arguments and can be used independently on every argument.

Dataset Creation

This section deals with the process followed for the creation of the dataset for argument quality analysis. We have broadly split this into three parts: Argument Collection, Argument Annotation and Argument Scoring.

Procedure for Argument Collection

Argument Collection for ArgAnalysis35K was primarily done through two ways.

1. A majority of argument-analysis pairs (~60%) were collected through contribution by a set of active debaters of varying levels of expertise. These people were recruited at debating tournaments, through active debate circuits, debating facebook groups and contacts of past/current debaters.
   - Experts: Won 5+ tournaments at a global or regional level or have 3+ years of active debating experience. Experts contributed around 22% of our argument-analysis pairs.
   - Intermediate: Won 2+ tournaments at a global or regional level or have 1-3 years of active debating experience. Intermediates contributed around 22% of our argument-analysis pairs.
   - Novice: Not won a tournament or < 1 year of debating experience. Novice debaters contributed around 15% of our argument-analysis pairs.
2. $\sim 40\%$ of argument-analysis pairs were extracted from speeches given in the outrounds of tournaments. The tournaments considered were regional majors (EUDC, ABP, UADC, etc.) or global majors (WUDC, WSDC). We also restricted the extraction to speeches given in the elimination stage (outrounds) of the tournaments, which is a good way to ensure a high quality of argument-analysis pairs. Only speeches from tournaments within the last 10 years (earliest 2011) were considered, both due to the need for maintaining relevance of arguments and the difficulty in finding good quality transcripts/videos of speeches given prior to that date.

While collecting arguments from contributors, we used the following procedure. Each contributor was presented with a single motion at a time and asked to contribute one argument for and one argument against the motion. It was explained that an argument is a statement in defence of or against the motion presented. Then, the contributor was asked to come up with analysis statements defending the arguments. An analysis statement was explained to be a reason why we find the specific argument persuasive. We also set a character limit of 20-210 for each argument and 35-400 for each analysis point. This limit was set taking into consideration that an argument is expected to be a mere statement that is short and impactful, and analysis is expected to have more content as it defends the argument. All argument contributions were on a non-compensated volunteer basis and the work-load for each volunteer was kept to a maximum of 20 minutes.

Argument Annotation Collection

A total of 200 individuals were involved in the annotation process for the dataset. The annotators chosen had participated in at least one debate at a school or college level. The experience level was set in order to better deal with the additional complexity of annotating argument-analysis pairs, since this concept is part of the fundamental training that is required to participate in a debate. They came from debating circuits all around the world to ensure that diversity (in arguments, thoughts, etc) is being expressed in the dataset. Considering the relatively high experience level of the annotators, each argument was annotated by three annotators. Each annotator was asked two questions per argument-analysis pair.

1. Is the argument something you would recommend a friend use as-is in a speech supporting/opposing a topic, regardless of personal opinion?
2. Would you recommend a friend use the analysis to defend the argument as it is?

The annotations were collected in six sessions over a period of four months. Each annotator was asked to annotate 100 arguments per session. Each session took approximately 120 mins. This meant that on average, each annotator spent more than a minute analysing an argument analysis pair, a time which was is sufficient to gain a representative understanding of how the annotator viewed the argument-analysis pair. In order to gauge whether an annotator was paying attention to the task, there was a hidden test question asking the annotator to leave the response field blank if they had read the question. Annotators that failed the hidden question twice were removed from the annotation process. Surprisingly for an endeavour of this size, only three annotators had to be removed for this reason (1.5% of the total pool).

Annotator Reliability Score and Tests

Annotator-Rel score is required for the calculation of the Weighted Average scoring function proposed by Gretz et al. (2020a). It is obtained by averaging all pair-wise κ for a given annotator, with other annotators that share at least 50 common judgments. Annotators who do not share at least 50 common judgments with at least 5 other annotators, do not receive a value for this score. The task-average κ is an important metric in this case to judge the overall quality of the annotation process. It is basically the average of all the pairwise-κ for all annotators. In comparison to Gretz et al. (2020a)'s reported value of 0.83, we find that our task-average κ value is 0.89. We hypothesise that this high value is due to the lower number of annotators involved and the comparatively higher and consistent experience level of the annotators. All annotation was done on a non-compensated volunteer basis.

Scoring Functions

Scoring an argument-analysis pair is an inherently subjective task. In order to make it as objective as possible, we have reduced the annotator involvement to two binary questions. However in order to make our dataset usable and interfaceable with others in the field (Gretz et al. 2020a; Habernal and Gurevych 2016b), we need to convert these annotations to a quality score. In order to do this, we have used the two methods used in the creation of IBM-30k as well as a third, recently proposed method that models annotator reliability on a per instance basis.

They were paid in compensation as well as arranged training sessions, personal debate coaching, competitions, etc as applicable in specific instances.

MACE-P

Since we have asked two questions, one related to argument and one to analysis, correspondingly, we have two scores generated per argument-analysis pair. We denote these scores as MACE-P_{Arg} and MACE-P_{Analysis}

Weighted Average

As mentioned previously, we utilize the annotator reliability we have calculated in order to compute Weighted Average scores for the two binary questions. As before, we get two scores per argument-analysis pair - WA_{arg} and WA_{analysis}

Instance-Based Annotator Reliability

We have applied a third scoring function to our dataset considering the following assumptions:

- Since we are selecting our annotators with a baseline level of expertise in the field of debating and have ruled out unattentive people, the remaining annotators are unlikely to be incompetent.
- Annotators are human and have human biases. They are likely to be biased, prejudiced and unreliable in specific instances

Considering these assumptions, we decided to apply the scoring function proposed by Li et al. (2019) as it seemed to be an ideal use case for their approach of modelling instance based annotator reliability. This method is basically a modified version of MACE and uses Expectation Maximisation training and FNN classifiers to generate per instance annotator reliabilities and use those to predict the true value of an annotation. In order to increase efficiency, we pretrain this model using a golden source annotation dataset. This dataset is created by sampling 500 collected argument-analysis pairs and getting them annotated by a set of 10 experts. Out of the 500 pairs, we observe 100% agreement between experts on 260 pairs. These 260 pairs form our golden source annotation dataset. Again, this method will generate two scores per pair - IA_{arg} and IA_{analysis}.

Aggregation of scores

Since we are scoring arguments and analysis separately, we have come up with two scores per scoring function discussed so far. Arguments and analysis are linked intrinsically in the context of debate. A good argument defended badly is non-persuasive, as is a bad argument defended well. In order to model this behaviour, we propose that to get the overall score of an argument analysis pair, we multiply the two scores together to get an overall score as shown in equation 1.

$$Score_{\text{pair}} = Score_{\text{arg}} * Score_{\text{analysis}} \tag{1}$$

These are people who have core adjudicated in multiple tournaments, made a name in renowned tournaments like WUDC/UADC/ABP and invited and paid to judge tournaments around the world. They were compensated appropriately for their respective contributions.

Scoring Function Comparison

Here, we have compared the three scoring functions described by performing the following experiments. Additional details about these experiments can be found in the appendix.

Disagreement in choosing the better argument-analysis pair

Here, we paired up argument-analysis pairs where we see a difference in scoring between MACE-P, WA and IA scoring functions. Annotators were asked to pick the argument-analysis pair that they would prefer to recommend to someone regardless of personal bias to use as-is. We then look at the agreement between the different annotators on each of the pairs. For those pairs differing in WA and IA, annotators preferred IA in 68% of the pairs. Similarly, for those pairs differing in IA and MACE-P, annotators preferred IA in 64% of the pairs.

Scoring Function	**Delta**	**Filtered Pairs**	**Precision**
WA_{pair}	< 0.25	11%	0.67
WA_{pair}	0.25-0.5	10%	0.72
WA_{pair}	0.5-0.75	8%	0.95
WA_{pair}	0.75+	4%	1.00
MACE-P_{pair}	< 0.25	11%	0.59
MACE-P_{pair}	0.25-0.5	10%	0.71
MACE-P_{pair}	0.5-0.75	8%	0.83
MACE-P_{pair}	0.75+	4%	0.90
IA_{pair}	< 0.25	11%	0.69
IA_{pair}	0.25-0.5	10%	0.73
IA_{pair}	0.5-0.75	8%	0.84
IA_{pair}	0.75+	4%	0.91

Table 3: Comparing Scoring Functions against Gold Standard Arguments, showing that the higher the delta between the scores, the higher is the precision value for annotators recognizing the higher rated pair.

Pairwise annotation Agreement

Another simple experiment that helps us determine the quality of the scoring functions is testing the agreement with pairwise gold-standard annotations. We place argument-analysis pairs in four bins as per the delta between the scores. The deltas used for the bins were as seen in table 3. From each of these bins, we created a random sample of 150 arguments and sent them for pairwise annotations just as in the last experiment. The same process was followed for all three scoring functions.

We find that MACE-P and IA tend to show similar precision for higher deltas but for lower bins, more annotators tend to agree with IA. This may be because of the additional nuance captured as a result of modelling annotator reliability on a per-instance basis

Scoring Function	Correlation Coefficient
$WA_{argument}$	0.74
$WA_{analysis}$	0.62
$MACE\text{-}P_{argument}$	0.69
$MACE\text{-}P_{analysis}$	0.60
$IA_{argument}$	0.70
$IA_{analysis}$	0.59

Table 4: Reproducibility Test Results

Reproducibility Test

Ideally, a scoring function should be consistent across the dataset. This means that if we were to sample the dataset and follow the same procedure of creating and scoring argument analysis pairs, we should end up with similar scores for the arguments. In order to perform this experiment, we randomly sample 500 argument-analysis pairs from our dataset and send them to a different set of annotators following the same procedure. We then calculate the Spearman's Rank Correlation Coefficient between the scores calculated using the new annotations and the scores calculated originally. We find that there is a strong correlation for all three scoring functions in terms of the argument scores, but that correlation gets slightly weaker when it comes to analysis scores. This can be explained due to the slightly more subjective nature of the analysis. In terms of the scoring functions, we find that there is a slightly higher correlation for weighted average as opposed to the other two methods, which is an observation that agrees with the previous experiment's findings.

Relevance Model

In this section, we describe the relevance model that quantifies the applicability of each argument-analysis pair to a topic. The underlying assumption is that each argument-analysis pair has a degree of applicability to at least one and likely more topics. This assumption is made on the basis of the personal experience that we have gathered while debating and discussions with experts in the field, where we often find that arguments repeat across multiple topics and motions. (Gretz et al. 2020b) conducted a qualitative evaluation of the correlation between relevance or applicability of an argument and a topic and how that is one of the factors by which we can understand why a particular argument is good. We believe that the approach can be extended in a quantitative manner by application of topic modeling and topic analysis.

Creation of the Relevance Model

In order to build our relevance model, we utilize the following algorithm.

1. We generate a list of 24 topics considering inputs from our experts, analysis of trends in debating and classification of motions that we had presented to our annotators in order to generate our arguments.
2. In order to get more nuance on these topics, we asked 50 annotators to come up with a list of 5 keywords (also referred to as subtopics) per topic. The annotators chosen for this task were the ones scoring the highest in the previous tasks we set.
3. The keywords were then aggregated for similarity and reduced to the simplest representation and the keywords with the most agreement between annotators ($< 70\%$ of annotators having included the keyword) were collected.
4. The list of keywords was then sent to the experts who were asked to classify them into two bins: one bin containing keywords that they perceived to be highly relevant to the topic and one bin containing keywords that they perceived to be not as relevant. The weight of the keyword was taken to be the percentage of experts placing the keyword in the high relevance bin.
5. The probability of each argument-analysis pair belonging to the topics was then calculated. This was achieved by applying W2V and BERT to generate a list of scores per argument-analysis pair and subtopic, which indicates the probability of the pair belonging to that topic.
6. These scores are then combined via the following formula to generate the overall relevance score of a particular argument-analysis pair to the main topic.

$$\frac{\sum_{i=1}^{n} \alpha_{percentage} * Prob_{BERT}}{\sum_{i=1}^{n} \alpha_{percentage}} \tag{2}$$

Preliminary Analysis of the model

We observe a small degree of overlap (approximately 15% of keywords having more than one non zero relevance score) in the keyword generation process, i.e. the same keyword being generated for different topics. We take this as evidence that there is a significant overlap of themes when it comes to debate. In this case they were assigned different weights for the different topics depending on the percentage of experts that placed the word in the high relevance bin for that particular topic. This created a set of 84 unique keywords with different weights for different topics.

Validation of relevance model

In order to validate the relevance model we propose a simple experiment. The hypothesis is that as the delta of relevance scores increases, it will be easier for annotators to identify which of the pair of arguments is more relevant to the given topic.

1. To make the comparisons fairer, we randomly select a topic for which the relevance scores will be considered.
2. We place argument-analysis pairs into four bins based on the delta of their relevance scores to the selected topic.
3. We then randomly sample 150 pairs and send them for pairwise annotations to a set of 50 people (highest scoring annotators and experts). Each annotator was asked to pick the more relevant argument for the given topic and the percentage of annotators picking the higher ranked argument was noted as the precision.

For the topic "Economy", the keywords "money", "rupee", "currency" all got reduced to money.

Topic	Delta	Filtered Pairs	Precision
Art	< 0.25	14%	0.72
Art	0.25-0.5	10%	0.77
Art	0.5-0.75	5%	0.84
Art	0.75+	2%	0.96

Table 5: Relevance Model Validation of the topic of art, similar analysis to be done for every topic

4. If sufficient agreement ($> 80\%$) between annotators was not achieved, the pair was dropped.

This procedure was followed for two more randomly sampled topics to ensure coverage of the dataset and the agreements with the relevance scores are recorded in Table 5. We found that all three topics showed similar trends in terms of agreeing with the annotator scoring. Annotator scoring also showed a high correlation with our relevance model for high deltas. This validates the relevance model as it satisfies the basic requirement of a quantitative score: bigger differences are more easily recognized.

Experimental Results

Experiments

We use several methods to learn the task of ranking the quality of arguments. We evaluate the following methods, some accepted standard baselines, some taken from Gretz et al. (2020a) and some other neural models.

- Arg Length: We evaluate the effect the length of an argument has on the scores of the argument to see if there is a correlation between the two, or if the annotators are biased to score longer arguments higher.
- Bi-LSTM GloVe: We implemented the model proposed by Levy et al. on a dropout of 0.10 and an LSTM layer of size 128. 300 dimensional GloVe embeddings were used for input features.
- BERT-FT$_{\text{topic}}$: Gretz et al. (2020a) has fine-tuned BERT to concatenate a topic parameter and replace the final softmax layer with a sigmoid function. This has achieved the best results for their dataset, hence for the purpose of comparison with a standard, we have tested our dataset through the same.

For the purpose of evaluating our methods on the ArgAnalysis35K dataset, we split the dataset into 70-20-10, 70% for training, 10% for tuning hyper parameters (to be used as a dev set), and 20% for testing. To keep the experiments consistent for comparing results with Gretz et al. (2020a), the same model parameters have been used: models have been trained for 5 epochs over the training data, with a batch size of 32 and a learning rate of 2e-5. Pearson and Spearman correlations are reported on the entire set.

Results and Discussion

The results are presented in Table 6. We find that argument length is not an indicator for quality. However, we notice an interesting trend when looking at analysis length with

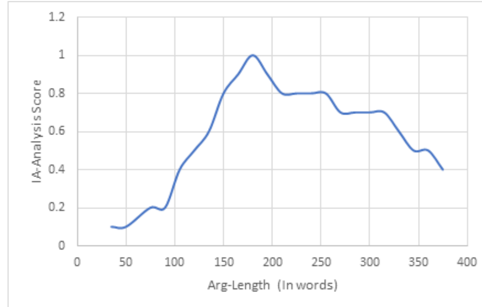

Figure 1: IA-Analysis Scores Vs Arg-Length, showing that argument quality score is not directly correlated to arg-length

comparison to the IA score they receive (Figure 1). Analysis scores reach a peak score at 180 characters, following which they drop, giving a slight resemblance to a normal curve. This proves that less characters are insufficient to express a point in a persuasive manner, but having more characters than necessary is also not considered persuasive, as the analysis becomes repetitive and less impactful. In order to compare the other scores effectively against existing datasets that do not have an analysis component, we aggregate the two scores per scoring function into one as described in section 4. ELMo improves on Bi-LSTM GloVe by 0.4-0.6 points as it is able to capture more nuance in the dataset, as opposed to GloVe. BERT-FT topic provides a significant improvement over the other methods.

Comparing Quality of ArgAnalysis35K Arguments to IBM-Rank30

Since WA has been used as a scoring function for ArgAnalysis35K as well as IBM-Rank30K, we are able to compare the scores of both datasets to compare argument quality. Out of the 5000 arguments ranked 1 in IBM-Rank30, we randomly sampled 200. We run these arguments through our relevance model to find the topic in our dataset they are closest related to. The specified argument was only taken if it had a relevance score above 0.8 (that is, the argument strongly belongs to that category). From the ArgAnalysis35K dataset, we have randomly selected an argument-analysis pair from the same topic that had been scored 1. This pair of arguments were then sent to 500 random debaters where they were asked which argument they found more persuasive (similar to the question asked during the debate between Project Debater and Harish Natarajan). We then look at the agreement between the different annotators on each of the pairs, similar to the experiment performed to compare the different scoring functions. We found that annotators preferred a ArgAnalysis35K argument 71% of the time, hence showing that the arguments in ArgAnalysis35K are more relevant in the context of parliamentary debating, and that an argument is more persuasive when followed by analysis.

Model	WA_{pair}		MACE-P_{pair}		IA_{pair}	
	r	ρ	r	ρ	r	ρ
Arg-Length	0.18	0.19	0.19	0.19	0.16	0.17
Analysis-Length	0.32	0.31	0.29	0.28	0.32	0.33
Bi-LSTM GLoVe	0.39	0.41	0.42	0.41	0.43	0.42
BERT FT $_{TOPIC}$	0.52	0.53	0.54	0.53	0.54	0.55

Table 6: Results for the scoring functions

Comparing the relative effect of argument and analysis for the overall score

One of the major purposes of asking annotators to answer two questions and reporting two separate scores of argument and analysis is to answer the question of what makes an argument persuasive: the argument itself or the explanation and analysis given for it. In order to test this, we plot a histogram of arguments and analysis separately against the distribution of the score (additional graphs attached in appendix). We find that analysis points have more scores above 0.7 than arguments alone, hence proving that logical links and explanations are critical to increase the persuasiveness of an argument.

Conclusion and Future Works

In this work, we create ArgAnalysis35K and validate it using a variety of mathods. This system can be integrated with existing models to create a system that is able to debate more efficiently, be more persuasive, and as a result win more debates.

Broader Impacts and Ethical Considerations

There are a few assumptions made in the process of creating the dataset. We have tried to cover the arguments involved in debating by talking to experts and people from debate circuits across the world, with different experiences and expertise. However, due to the nature of this activity, it is possible that there are arguments and experiences have not been covered in the dataset. These could be experiences of marginalized communities, underrepresented debate circuits, etc. We also assume that the 24 topics are indicative of all topics in debating. We have validated our list using data from multiple tournaments, experts, Core adjudicators to ensure that the maximum possible amount of diversity is incorporated. We have included a large number of high quality arguments, unlike other similar projects, to increase the possibility of creating a system capable of winning against a human, a chance that is otherwise missing with other datasets. The number of annotators used to create and validate the dataset and its functions is small (200 at most), we find that this is on par with similar projects. We have compensated all annotators as applicable. Lastly, even though arguments were taken from WUDC speeches by watching and recording them, they were anonymized by removing names, paraphasing the argument and making it otherwise unrecognizable to point out where an argument came from (even for an expert debater).

References

Bazari, S.; Leader Maynard, J.; Arıkan, E. F.; Madeline Schultz, B.; Templeton, S.; van Koppenhagen, D.; Baer, M.; Block, S.; Cochrane, D.; David, L.; Natarajan, H.; Parmanand, S.; Li, S.; Tuffin, A.; Roussos, J.; Dobranić, F.; Kirova, D.; and Nevo, O. 2015. The Worlds University Debating Championship: Debating and Judging Manual.

Ein-Dor, L.; Shnarch, E.; Dankin, L.; Halfon, A.; Sznajder, B.; Gera, A.; Alzate, C.; Gleize, M.; Choshen, L.; Hou, Y.; Bilu, Y.; Aharonov, R.; and Slonim, N. 2019. Corpus Wide Argument Mining – a Working Solution. arXiv:1911.10763.

Gretz, S.; Friedman, R.; Cohen-Karlik, E.; Toledo, A.; Lahav, D.; Aharonov, R.; and Slonim, N. 2020a. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis. In *AAAI*.

Gretz, S.; Friedman, R.; Cohen-Karlik, E.; Toledo, A.; Lahav, D.; Aharonov, R.; and Slonim, N. 2020b. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7805–7813.

Habernal, I.; and Gurevych, I. 2016a. 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*, 1214–1223. Austin, Texas: Association for Computational Linguistics.

Habernal, I.; and Gurevych, I. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1589–1599. Berlin, Germany: Association for Computational Linguistics.

Hua, X.; and Wang, L. 2017. Understanding and Detecting Supporting Arguments of Diverse Types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 203–208. Vancouver, Canada: Association for Computational Linguistics.

Li, M.; Fahlström Myrman, A.; Mu, T.; and Ananiadou, S. 2019. Modelling Instance-Level Annotator Reliability for Natural Language Labelling Tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2873–2883. Minneapolis, Minnesota: Association for Computational Linguistics.

Paglieri, F.; and Castelfranchi, C. 2014. Trust, relevance, and arguments. *Argument Computation*, 5.

Persing, I.; and Ng, V. 2017. Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4082–4088.

Roush, A.; and Balaji, A. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, 1–7. Online: Association for Computational Linguistics.

Slonim, N.; Bilu, Y.; Alzate, C.; Bar-Haim, R.; Bogin, B.; Bonin, F.; Choshen, L.; Cohen-Karlik, E.; Dankin, L.; Edelstein, L.; Ein-Dor, L.; Friedman-Melamed, R.; Gavron, A.; Gera, A.; Gleize, M.; Gretz, S.; Gutfreund, D.; Halfon, A.; Hershcovich, D.; Hoory, R.; Hou, Y.; Hummel, S.; Jacovi, M.; Jochim, C.; Kantor, Y.; Katz, Y.; Konopnicki, D.; Kons, Z.; Kotlerman, L.; Krieger, D.; Lahav, D.; Lavee, T.; Levy, R.; Liberman, N.; Mass, Y.; Menczel, A.; Mirkin, S.; Moshkowich, G.; Ofek-Koifman, S.; Orbach, M.; Rabinovich, E.; Rinott, R.; Shechtman, S.; Sheinwald, D.; Shnarch, E.; Shnayderman, I.; Soffer, A.; Spector, A.; Sznajder, B.; Toledo, A.; Toledo-Ronen, O.; Venezian, E.; and Aharonov, R. 2021. An autonomous debating system. *Nature*, 591(7850): 379–384.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Toledo, A.; Gretz, S.; Cohen-Karlik, E.; Friedman, R.; Venezian, E.; Lahav, D.; Jacovi, M.; Aharonov, R.; and Slonim, N. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5625–5635. Hong Kong, China: Association for Computational Linguistics.

Toledo-Ronen, O.; Bar-Haim, R.; Halfon, A.; Jochim, C.; Menczel, A.; Aharonov, R.; and Slonim, N. 2018. Learning Sentiment Composition from Sentiment Lexicons. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2230–2241. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Wachsmuth, H.; Potthast, M.; Al-Khatib, K.; Ajjour, Y.; Puschmann, J.; Qu, J.; Dorsch, J.; Morari, V.; Bevendorff, J.; and Stein, B. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, 49–59. Copenhagen, Denmark: Association for Computational Linguistics.

Wachsmuth, H.; Stein, B.; and Ajjour, Y. 2017. "PageRank" for Argument Relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1117–1127. Valencia, Spain: Association for Computational Linguistics.

Appendix

Additional Details: Disagreement in choosing the better pair (5.1)

The argument-analysis pairs chosen in this experiment belonged to the same stance on the same topic, in order to avoid annotator bias. This generated a dataset of 737 pairs. The dataset was then split between a set of individuals comprising the highest scoring annotators and experts (around 50 individuals). Each argument was seen by 5 individual annotators and this annotation was done in a single session. IBM-30k used a threshold of 70% agreement between annotators to pick out the final set of pairs in their experiment. Since we used a high threshold to select annotators for this task, we set a correspondingly higher threshold of 80% agreement between all annotators to drop the pairs.

This results in a similar percentage of pairs being dropped (28%) and we are left with a total of 530 pairs. Out of them, 368 are differently ranked for MACE-P and WA, 250 are differently ranked for WA and IA, and 90 are differently ranked for MACE-P and IA. A reason for this disparity might be the relatively similar methodologies followed by MACE and IA.

Additional Details: Pairwise Annotation Agreement (5.2)

The assumption here is that pairs with a higher delta should show a higher agreement with annotations as it should be easier for annotators to identify the better argument-analysis pair in case of a huge difference in quality. In order to test the agreement with this assumption, we tabulate the results of precision against delta for the three scoring functions. We drop the pairs that do not show sufficient agreement between annotators, a threshold that we set at 80% due to the reasons mentioned above. The results we record for the comparison between MACE-P and WA agree with the ones reported by Gretz et al. (2020a). We find that considering the pairs with delta more than 0.25, that precision tends to be better for WA than either of IA or MACE-P.

Additional Details: Reproducibility Test (5.3)

In this experiment, we did not combine the argument and analysis scores to generate a single score for the pair, as we wanted to gauge the effect of re-scoring the dataset on each of the individual components of our scores and scoring functions.

Scoring Function Additional Details

Overall, we believe that all three of the scoring functions evaluated in this section have unique value when it comes to highlighting different aspects of the dataset. Overall we observe a higher proportion of extreme values for both Weighted Average and MACE-P functions. This might be because of the context lost by dropping all annotator scores below a certain threshold making the resulting annotations more homogeneous. IA on the other hand, tends to provide a much smoother curve as we attempt to preserve as much contribution from each annotator as possible, thus leading to a more representative annotation set. Furthermore, Weighted Average tends to generate a continuous scoring scale while

MACE-P tends to cluster argument-analysis pairs around either of the two extremes, but we observe that IA offers a middle ground approach to get as close to the true value of an argument as possible, while still maintaining a smooth, continuous scoring curve. However, in order to make our dataset interfaceable with others in the field and to not lose out on the value generated by the other two scoring functions, we report all six scores in the final dataset.

Source	Number of Arguments	$MACE_{Arg}$ Average	$MACE_{Analysis}$ Average	WA_{Arg} Average	$WA_{Analysis}$ Average	IA_{Arg} Average	$IA_{Analysis}$ Average
WUDC Speech	13995	0.76	0.93	0.75	0.91	0.77	0.93
Expert Debater	7852	0.81	0.95	0.78	0.92	0.80	0.94
Intermediate Debater	7796	0.69	0.87	0.69	0.86	0.70	0.88
Novice Debater	5247	0.56	0.66	0.53	0.63	0.55	0.65
Total	34890	0.73	0.88	0.71	0.86	0.73	0.89

Table 7: An overview of the different sources of arguments and corresponding scores. The total number of rows in the dataset = 34980 * 24 scores for relevance = 839,520.

Argument	Analysis	IA_{Arg}	$IA_{Analysis}$	Score
Monopolies can justify spending money on R&D which smaller companies cannot do, and hence it is okay to keep a monopoly like Facebook running in the modern day.	Monopolies do not have competition and hence they are not worried about other companies taking over, which is why they can justify the risk of spending money on R&D which might or might not work.	1	1	WUDC Speech
Big companies are bad.	Since markets are a zero sum game, billionaires and big companies are not benevolent; they have stepped on others and exploited workers, customers to get there.	0.12	0.93	Intermediate Debater
Prioritizing being a monopoly over short term profit leads to an Increased power disparity between companies and consumers.	Customers are a vulnerable target.	0.81	0.22	Novice Debater

Table 8: An example of argument-analysis pairs from different sources with IA scores

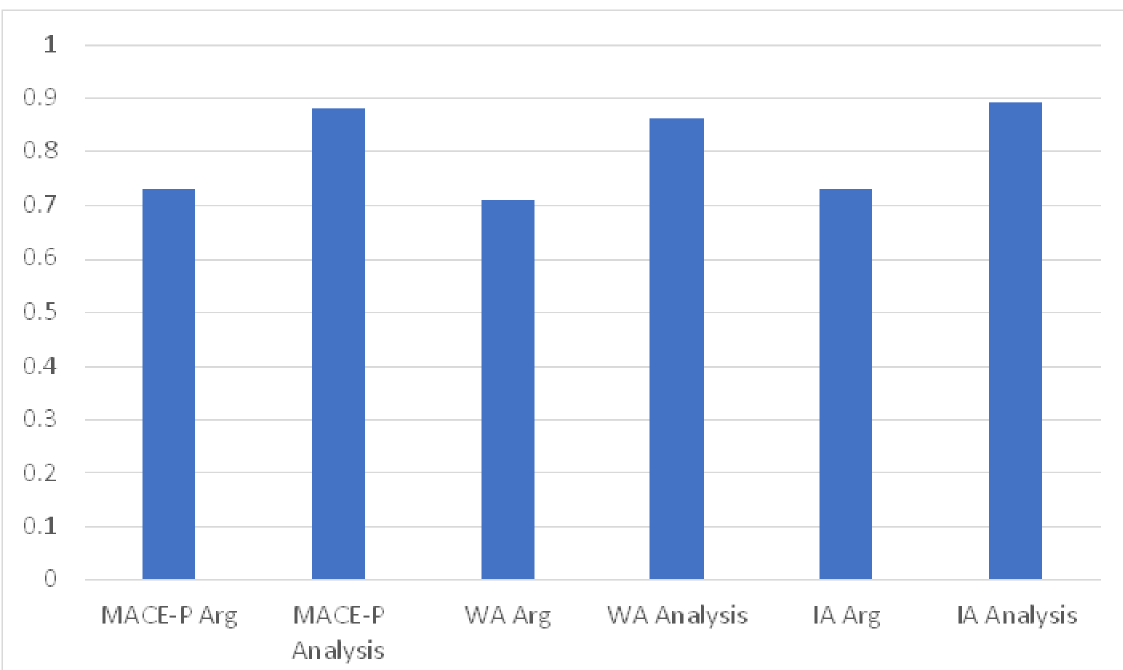

Figure 2: Scoring Functions, showing the importance of including analysis, which has a higher score

Topic	**Keywords**
Authoritarian Regimes	Russia, Dictatorship, China
Politics	Elections, Democracy, Vote
Diplomacy	International Relations, Negotiations, Foreign Policy
Economics	Cryptocurrency, Recession, Fiscal deficit
Philosophy	Nihlism, Rationalism, Stoicism
Morality and Ethics	Consent, Principles, Parenting
Criminal Justice	Punishment, Rehab, Juries
Social Justice	Discrimination, Racism, Philanthropy
Collective Action	Feminism, LGBTQ, Racism
Education	Syllabus, Teachers, Privilege
Art and Culture	Heritage, History, Commercialization
Business	Taxes, Facebook, Banks
Developing Nations	Post-colonialism, Pollution, Overpopulation
Environment	Climate Change, Pollution, Philanthropy
Family and Relationships	Parenting, Marriage, Toxic
Media	Social Media, Polarization, Depression
Religion	Atheism, Separation of powers, Divinity
Science and Technology	AI, Patents, Medicines
War and Terrorism	Drones, Decapitation, Death penalty
Sports	Children, Cult of personality, Leagues
Human Experience	Pessimism, Optimism, Death
Policy	Government, Whistleblowers, Immigration
International Organizations	UN, NATO, WTO
Diseases and Medicine	Pandemic, Therapy, Big pharma

Table 9: A list of topics and chosen keywords