The generalizability of argument quality dimensions in NLP models

Thesis report

by

Jakub Nguyen

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on August 15, 2023 at 11:00

Student number: 4904540
Thesis committee:

Dr. Catholijn Jonker	Thesis advisor
Dr. Pradeep Murukannaiah	Daily supervisor
Michiel van der Meer	Daily co-supervisor
Dr. Jie Yang	External examiner

CONTENTS

SUMMARY

This research revolves around measuring the quality of arguments. High-quality arguments help in improving political discussions, resulting in better decision-making. Wachsmuth et al. developed a taxonomy breaking down argument quality into several dimensions. This work makes use of that taxonomy and combines it with modern NLP models. A cross-dataset examination of argument quality models was conducted. In particular, models were investigated on their generalizability between dimensions. Overall results show that there is no large difference in accuracy and agreement when models predict data of a quality dimension they were trained on, over dimensions they were not trained on. One can conclude that generalizations of argument quality dimensions with language models were not found. Nevertheless, qualitative analysis highlights findings that indicate some generalization to other dimensions.

1

INTRODUCTION

1

Striving to make good political decisions is an essential aspect of deliberative democracy [1]. Deliberative democracy is part of political theory and states that political decisions should originate from rational and fair debates. Such debates consist of arguments presented by opposing parties. Since these debates can lead to political decisions they are of great importance to society. However, such discourses do not only take place in parliaments but also among citizens in local communities. Consider citizen participation [2] where laypeople may contribute to decision-making on a municipal level.

Improving upon these discussions would be of great value, especially if automated. One could imagine a form of moderation that provides additional information to each argument made. Nevertheless before improving such debates, one needs to understand how to assess arguments first. The assessment of arguments is not a trivial task but the field revolving around *computational argumentation* takes on this challenge.

There are multiple reasons for the difficulty of assessing the quality of arguments.

At its core argument quality is a subjective concept since there can be great differences in values, beliefs, and personal experiences among people. For instance, some may perceive an argument as disrespectful while others do not.

In order to evaluate arguments appropriately it is also important to consider the context. Measuring the quality of arguments may be very different when looking at political debates or classroom discussions. Furthermore, this relates to the complexity of linguistics as it is not always simple to understand irony, sarcasm, and figurative language even as a human.

Beyond helping in the moderation of debates, argument quality estimation can be used to assist in writing [3] or searching new arguments [4]. It also allows for more informed decision-making, identifying weaknesses in argumentation, and evaluating the credibility of claims.

There exists a plethora of research that aims to assess argument quality in some way, see [5] [6] [7] [8]. Due to the vast amount of options to assess argument quality, this research focuses on the work by Wachsmuth et al. [9] since it is one of the most referenced ones [10] [11]. Wachsmuth et al. propose a framework that lays the foundation for a common understanding to assess arguments on a more granular level. With that framework argument quality can be split up into sub-dimensions. The research shows that there exist positive correlations between these sub-dimensions. This research aims to empirically investigate the relations between some of these dimensions.

Nowadays there are various data sets available which depict dimensions of argument quality. Usually, such data is gathered from different contexts such that one might not obtain the same understanding of one dimension from multiple data sets. For instance, stating that an argument is sufficiently supported is quite different when being in a parliamentary debate or classroom discussion. This also goes the other way, it may be possible that data sets portraying different dimensions could be expressing a similar concept. For instance, when considering dimensions like persuasiveness and convincingness, there indeed exists a conceptual difference. On the other hand, it may be quite difficult to determine whether an argument is more convincing than it is persuasive in practice. Particularly when considering means like crowd-sourcing to obtain annotations nowadays.

The possibility of generalizing dimensions can be expanded to the domain of computational argument quality where NLP users make use of the aforementioned data sets. NLP models that are obtained by training on such data may be able to correctly predict argument quality dimensions other than the one trained on. Wachsmuth et al. [9] indicate relations between dimensions in Figure 1.1 of which selected will be investigated in this research. We introduce the term of *cross-prediction* which depicts label predictions on one data set performed by a model trained on another data set.

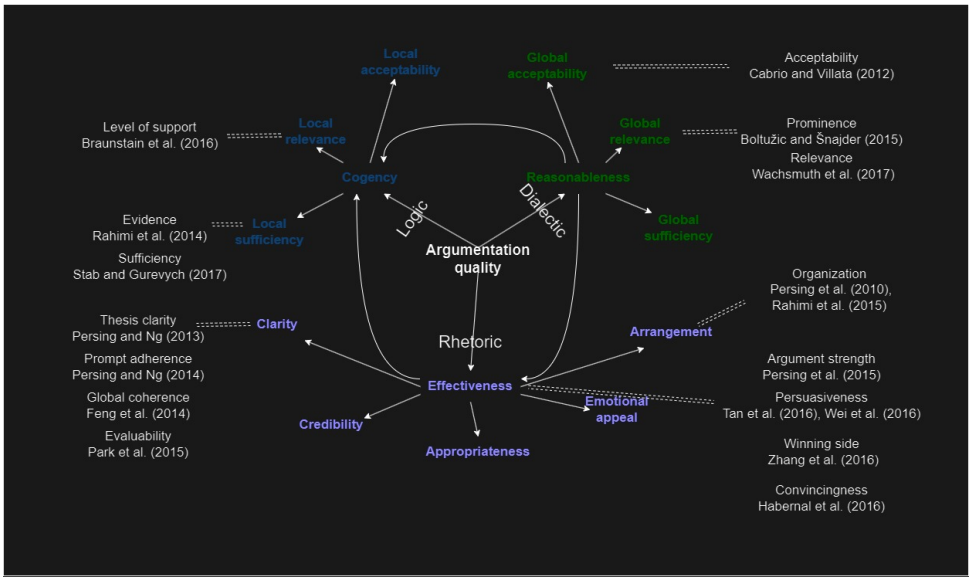

Figure 1.1: The proposed taxonomy of argumentation quality as well as the mapping of existing assessment approaches to the covered quality dimensions. Arrows show main dependencies between the dimensions. Image taken from Wachsmuth et al. 2017 [9]

To investigate the generalization of dimensions in NLP models this research aims to contribute by answering the following questions:

RQ1 *Do argument quality prediction models, trained on one quality dimension, generalize from one data set to another (based on the taxonomy by Wachsmuth et al.)?*

RQ1.1 *Are cross-predictions within dimensions more accurate than cross-predictions across dimensions?*

RQ1.2 *Does changing the base model affect the generalizability?*

By observing some generalizability of argument quality dimensions one can grasp relations between dimensions. This could allow for verifying and complementing existing argumentation theories. Furthermore, it gives insights into how well NLP models can learn argument quality estimation. Contrary, if no generalizability can be found it hints toward larger differences in dimensions or data sets.

1

The thesis is structured in the following manner. First, we inspect related work in the domains of argument quality prediction and transfer learning. Then the method to answer the presented research questions is described. According to the method the experimental setup is presented followed by the results. Lastly, the discussion of results and conclusion are covered.

2

Related Work

2.1. Argument Quality estimation

Previous research in the field of argument quality estimation provided valuable insights into its computational prediction. For one there exists work that focuses on Arguments as a whole. Gretz et al. [12] devised a large annotated dataset of arguments through crowd-sourcing. An argument with its topic was provided to the annotator who then needed to decide whether they would recommend using that argument in a fitting speech. This aimed at conceiving the overall quality of the argument. By utilizing multiple annotations for each argument Gretz et al. ranked the arguments based on two scoring functions. The annotated dataset was then used to train machine-learning models including fine-tuning BERT base. It was shown that neural learning approaches significantly outperformed the baseline ranking methods.

There also exist lines of research dedicated to specific quality aspects of arguments. Rahimi et al. [13] engaged in the *evidence* of essays. They investigated how well an essay supports its claims and used three learning methods to evaluate their effectiveness to predict the *evidence*. The devised models were based on Naive Bayes, Random Forest and Logistic Regression. It was found that essays that relate more often and explicitly to their problem statement support the prediction process significantly. In a similar fashion Stab and Gurevych [14] investigated the *sufficiency* of arguments, i.e. if the premise of an argument can be deduced from supporting statements the argument is sufficient. Based on the results of their experiments it was shown that convolutional neural networks outperform feature-based classification.

Beyond the aspects of logic in argumentation, it is also possible to delve into the rhetorical components of arguments. Tan et al. [15] utilized the public data from a Reddit community called "change my view". They show that the persuasiveness of an argument is positively affected by the number of interactions between participants as well as the reputation of the argument's author. Persing at al. take a different approach by inspecting the arrangement of essays [6]. They demonstrate various techniques using the essay's arrangement to obtain a score for each essay.

Wachsmuth et al. presented a taxonomy that allows for breaking down argument quality into multiple dimensions [9]. It is based on existing argument quality theories, assembling them in one overview. The quality dimensions can be categorized by the aspects of logic, rhetoric, and dialect. Furthermore, they devised a dataset that is annotated corresponding to their taxonomy.

In another study by Lauscher et al. [16] the researchers followed the framework proposed by Wachsmuth (2017). They focused on the three main aspects of argumentation quality: Cogency, Effectiveness, Reasonableness, and overall Argument Quality (AQ). The researchers created their own dataset which focuses on the differences in data domains such as Yahoo! Answers, Reddit Change my view, and Yelp online reviews. The dataset was annotated by experts and through crowd-sourcing for each of the three AQ aspects. To gather annotations, they formulated questions related to subdimensions and presented them to annotators. Each aspect was labeled on a scale of 1 to 5. Various models were trained on this dataset, including three BERT models using linear regression, with each model dedicated to one aspect. Additionally, two forms of multi-task learning were employed, training on all dimensions and the overall AQ. The models were then used to predict different data sets grouped by their respective domains. Although they briefly investigate the effect of interrelations between AQ dimensions, this aspect was not the primary focus of their research. The researchers conducted experiments with IBM-Rank-30k, referred to as IBM argQ30k in this study, training their dataset and predicting the other dataset, and vice versa. This investigation aimed to assess the relationship between practical and theory-based argument quality assessment. When training on their data set, GAQCorpus, and predicting on IBM-Rank-30k, predictions for the dimension of *effectiveness* have shown to have the highest correlation with IBM-Rank-30k data. Also the model trained on arguments from the domain of debate forums achieved the highest correlation with the IBM-Rank-30k data. This is contrary to expectations since one model was trained on all domains. When training on the IBM-Rank-30k data and predicting their newly devised data a significant loss in performance was measured.

The IBM argQ30k data set was devised by Gretz et al. [12]. It was created through a crowd annotation task and partially expert annotations. Participants were asked to provide an argument for and against a given controversial topic. To obtain labels for these arguments another crowd-sourcing task was conducted. Workers were asked whether they would recommend using a given argument, irrespective of their personal views. With multiple labels for each argument, two scoring functions were used to obtain continuous scores. Beyond argument quality as a whole Gretz et al. also investigated argument quality dimensions based on the framework by Wachsmuth et al. [9]. They once again conducted a crowd-sourcing task in which workers were asked to label arguments for 10 dimensions. The findings were that arguments with a high-quality score, obtained from the previous tasks, also received above-average labels for each dimension. The same follows for low-quality arguments which obtained below average ratings for each dimension.

2.2. Transfer learning for NLP

Transfer learning in natural language processing (NLP) refers to a technique where a language model, trained on a data set, is used as a starting point for solving another

NLP task. Instead of training a model from scratch for a specific task, transfer learning allows us to utilize the knowledge learned by the pre-trained model, which often shows to improve performance and reduce the amount of labelled data required for training [17].

The key idea behind transfer learning in NLP is that the knowledge acquired by a model while learning one task can be useful for learning other related tasks [18]. This is based on the assumption that there are shared structures or semantics across different NLP tasks. By making use of the pre-trained model, one can effectively transfer this knowledge from the source data set to the target task.

The most common approach to transfer learning in NLP involves using a pre-trained language model, such as BERT or GPT. These models are typically trained on large-scale datasets to learn general language representations. They capture rich contextual information and semantic relationships between words, which can be valuable for various NLP tasks.

Transfer learning consists of pre-training and fine-tuning. First, a language model is trained on a large corpus of text data using unsupervised learning. The model learns to predict missing words in sentences, classify sentence relationships, or perform other auxiliary tasks that encourage it to understand the context and meaning of the text. This pre-training phase allows the model to learn general language representations that capture useful information about words, phrases, and sentence structures.

After pre-training, the pre-trained model is further trained on a task-specific dataset, which is labeled to match the target task. This fine-tuning step adapts the pre-trained model to the specific task by updating the model's parameters based on the labeled data. The pre-trained model's knowledge serves as a strong starting point, and the model can quickly learn the task-specific nuances by training on the smaller dataset. Fine-tuning typically involves adjusting the final layers of the model and sometimes the earlier layers as well.

By making use of transfer learning, one profits from the pre-trained model's ability to understand language and its contextual representations. This approach can greatly improve performance, especially when labeled data for a specific task is limited. Transfer learning has been successfully applied to numerous NLP tasks, such as sentiment analysis, named entity recognition, question answering, text classification, and machine translation, among others.

There have also been advancements when using transfer learning in the domain of argument mining. Wambsganss et al. [19] have already devised a model and pipeline to standardize the transfer learning process in argument mining. Using their method, they show on multiple occasions that their training results in state-of-the-art performance. This work lays the foundation for access to transfer learning without in depth domain knowledge in argument mining.

The research by Hua and Wang makes use of transfer learning to extract the structure of arguments [20]. They argue that the understanding of argument structures is similar across domains, making it transferable. This opens up possibilities to make use of more data, overcoming the scarcity of domain-specific argument structure annotations. Results show that not every domain benefits from transfer learning, hinting toward a unique language style, argumentative structure or the model not properly understand-

ing the argumentation. Nevertheless, their approach shows consistent success in other domains throughout extensive testing.

3

METHOD

To investigate the posed research questions the taxonomy of argument quality devised by Wachsmuth et al. is used as a starting point. In order to answer the research questions one needs data to train language models in the first place. In particular, data that fits the taxonomy and is publicly accessible. Furthermore, it should be compatible with the data set that was created based on the taxonomy, the *Dagstuhl-15512 ArgQuality Corpus*. Wachsmuth et al. outline several existing approaches to automatic argument quality assessment corresponding to the taxonomy making it a great starting point to find matching data sets. Beyond that, the Webis group gathered a plethora data sets on their webpage[1]. Since those stem from the domains of information retrieval, natural language processing data mining, and machine learning they form another search space in this research. Lastly, the IBM Project Debater data sets[2] offer another scope to explore. They were developed along the way of creating an automatic debating system, fitting into the field of argument mining and a search area for this research.

To then investigate the research questions a state-of-the-art NLP model needs to be selected as a base model to be used for the evaluation. For each data set found, that meets the aforementioned requirements, an instance of the base model is then fine-tuned on the data. Over the course of several epochs the model will be trained and refined. Eventually the model with the best performance on the validation set that has been encountered during training will be selected. The performance of each of these fine-tuned models is then measured through the predictions made. Predictions are performed on the same set of data sets, depicting certain quality dimensions. Naturally, if the model was already trained on a data set it was only used to predict the validation part of that data set. The assumption is that if two quality dimensions are generalizable by NLP models, models that are trained on one of the dimensions are able to predict data sets of the other dimension reliably. Furthermore, as a verification, we intend on inspecting at least two data sets per quality dimension to also have a measure for the generalizability of one data set to another within the same quality dimension.

[1] https://webis.de/data.html
[2] https://research.ibm.com/haifa/dept/vst/debating_data.shtml

In order to answer research question 1.2 multiple base models are needed. Performing more experimentation with those would lead to more variation and certainty in results. Each model is then once again fine-tuned on every data set, individually, from the previous research question. The fine-tuning procedure is standardized, aiming for a fair comparison of fine-tuned models. These fine-tuned models are then all tested to predict the quality dimension of every dataset obtaining information to gain insights into the generalizability of quality dimensions.

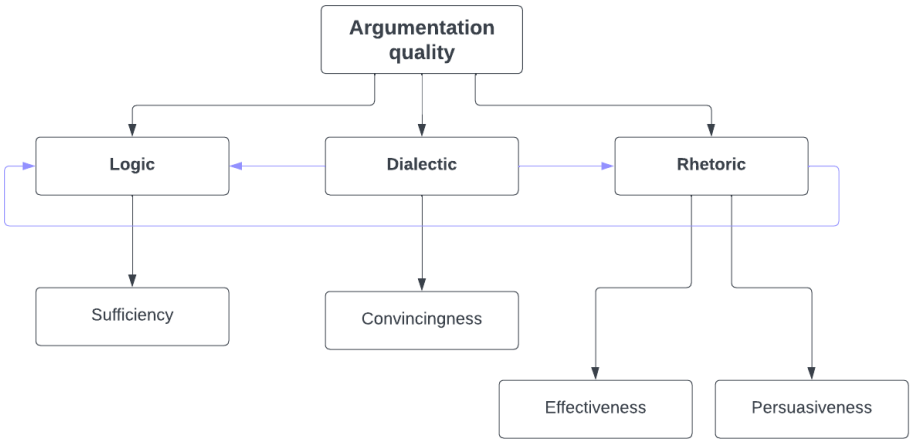

Figure 3.1: The taxonomy devised by Wachsmuth et al. includes only the quality dimensions researched in this work.

3.1. DATA

For a data set to be selected in this research, it needs to be annotated with one of the dimensions from the taxonomy of argumentation quality. Furthermore, the data set should be publicly accessible, that is the procedure of how the data set was created is well documented, the data set can be downloaded and it entails the raw argumentation text. As already pointed out in the method we use the *Dagstuhl-15512 ArgQuality Corpus* as a starting point since it directly corresponds to the taxonomy by Wachsmuth et al. and meets all other requirements. Since we selected a first data set it is now possible to define compatibility requirements for further selection of data sets. To retain a data set's meaning it is important to keep its original format and structure if possible. Nevertheless, when comparing data sets, through model predictions, some adaptations are needed unless all data sets have the same format. Therefore we focused on the most common formats found, being 1. an argument with a binary label describing if a quality dimension is fulfilled and 2. two arguments and a binary label, selecting one of the two arguments. Examples and further details can be found in the task description section.

Beyond the aforementioned requirements, a data set needs to meet to be selected for this study we also aim to depict the taxonomy in a fair way. Every category of quality

Aspect	Quality dimension	Source	Size	Status
Logic	**Sufficiency**	Stab and Gurevych (2017)	1029 arguments	accessible and compatible
	Evidence	Rahimi et al. (2014)		not accessible / not found
	Level of support	Braunstain et al. (2016)	50 advice-seeking questions	accessible but not compatible
Rhetoric	Argument strength	Persing and Ng (2015)	830 essays	not accessible, must purchase raw text data
	Persuasiveness	Tan et al. (2016), Wei et al. (2016)	3456 pairs	accessible and compatible
	Effectiveness	Al-Khatib et al. (2020)	10,303 pairs	accessible and compatible
	Prompt adherence	Persing and Ng (2014)	830 essays	not accessible, must purchase raw text data
	Thesis clarity	Persing and Ng (2013)	830 essays	not accessible, must purchase raw text data
	Winning side	Zhang et al. (2016)	108 debates	accessible but not compatible
Dialectic	Acceptability	Cabrio and Villata (2012)	219 arguments	not accessible, broken link
	Convincingness	Habernal and Gurevych (2016)	1052 arguments	accessible and compatible
	Prominence	Boltuzic and Snajder (2015)	3104 sentences	accessible but not compatible
	Relevance	Wachsmuth et al. (2017)	111 premise conclusion pairs	accessible but not compatible

Table 3.1: The findings of exploring the approaches to automatic argument quality assessment described by Wachsmuth et al.

Argument quality dimension	Name	Size	Annotators
Logic - Sufficiency	UKP: Insufficiently Supported Arguments in Argumentative Essays	1029 arguments	3
Several quality dimensions among them: Logic - Sufficiency Rhetoric - Effectiveness	Dagstuhl	320 arguments	3
Dialect - Convincingness	ACL2016 Convincing Arguments	11,652 argument pairs	1
Dialect - Convincingness	IBM Debater Dataset	5,698 argument pairs	1 (aggregated)
Rhetoric - Effectiveness	Webis ChangeMyView Corpus 2020	10,303 argument pairs	1 (post creator)
Rhetoric - Persuasiveness	Cornell ChangeMyView Data v1.0	3,456 argument pairs	1 (post creator)

Table 3.2: Data sets used for experiments

dimensions, logic, rhetoric, and dialectic needs to be represented in the selection of data sets. Furthermore, at least two data sets are selected for one dimension.

In the following, the findings of exploring the approaches to automatic argument quality assessment described by Wachsmuth et al. are shown. The quality dimensions written in bold font were selected for further experiments. For the dimensions of Evidence and Acceptability the authors of the respective research were contacted through email, attempting to obtain access to the data sets.

In Table 3.2 one can see which data sets were eventually selected as well as their format and number of annotators. These include the ones from the search through the Webis and IBM Debater data sets.

3.2. TASK DESCRIPTION

Since data sets are provided in different formats we choose two tasks such that minimal restructuring of data is required for training. Keeping the depiction of each quality dimension as much as possible.

Task 1 We already introduced different dimensions for measuring argument quality. We train models for rating argument quality per dimension. Instead of making a judgment per argument, we compare arguments pairwise, and pick the better argument. If available the corresponding topic is prepended to each argument.

Example

A0 We should adopt vegetarianism. Nicholas Stern, the author of the 2006 Stern Review on climate change has stated "people will need to turn vegetarian if the world is to conquer climate change".

A1 "We should adopt vegetarianism. Schopenhauer's views on animal rights stopped short of advocating vegetarianism, arguing that, so long as an animal's death was quick, men would suffer more by not eating meat than animals would suffer by being eaten."

Label: 1 (Argument 1 or A1 is sufficient)

Task 2 Complementary this task only considers individual arguments and requires a binary classification of the respective quality dimensions. Similarly, the topic is added to the argument if available.

Example

A0 It is also worth mentioning that some harmful effects on our health are lethal. It has been proved that overusing of the electronic devices including mobile phones could lead to higher possibility of suffering hearing loss and even cancers, although the further investigation are needed.

Label: 1 (the argument is sufficient)

4

Experimental Setup

To investigate the possible generalization of quality dimensions of argument quality we train NLP models to predict the dimension presented earlier. Fine-tuning is performed on 3 different base models. Namely *bert-base-uncased*, *ibm/ColD-Fusion-bert-base-uncased* and *IBM ArgQ30k-bert-base-uncased*. This fine-tuning was conducted with the same settings for each model, a standardized pipeline. The overall choice of models follows the study by Choshen et al. [21]. They investigated the effect of intermediate training of models on a prediction task. The terminology of source and target data set is used where the source data set consists of the data a model is fine-tuned on and the target data set is the data that is used for the prediction task. The selection of models follows the idea of Choshen et al. where the impact of the alignment between the source and target data set was investigated. To inspect the ability of NLP models to generalize from one argument quality dimension data set to another, we use *bert-base-uncased* as the base model. To then assess the effect of different base models for such generalization we consider *ibm/ColD-Fusion-bert-base-uncased* and *IBM ArgQ30k-bert-base-uncased*. ColD-Fusions source data sets have poor to no alignment with the target data set, therefore it was not trained on explicit argument data. Nevertheless, it was evaluated by Choshen et al. to score well across many different tasks, resulting in having the largest gain with bert-base-uncased as its base model. On the other hand, there is the *IBM ArgQ30k-bert-base-uncased* model which was fine-tuned on arguments that are annotated for argument quality in general, a good alignment of source and target data sets.

To assess the performance of the trained models the accuracy to correctly predict quality dimensions is used. Furthermore, to also inspect if models do similar mistakes the Cohens kappa score [22] is used, depicting the agreement between the predictions of models. It is defined as follows:

$$k = \frac{p_o - p_e}{1 - p_e}$$

Where p_o describes the relative observed agreement among raters and p_e depicts the hypothetical probability of agreement by chance.

Through the aforementioned setup, we can identify each prediction from the experiments with the following variables:

- **base model** bert-base-uncased / Cold-fusion / IBM ArgQ30k
- **source data set** an argument quality dimension data set to **train** on
- **target data set** an argument quality dimension data set to **predict**

In total we have three possible **base models** and seven possible **source data sets** resulting in 21 fine-tuned models. Each of these models can perform prediction on each of the seven possible **target data sets** resulting in 147 accuracy predictions.

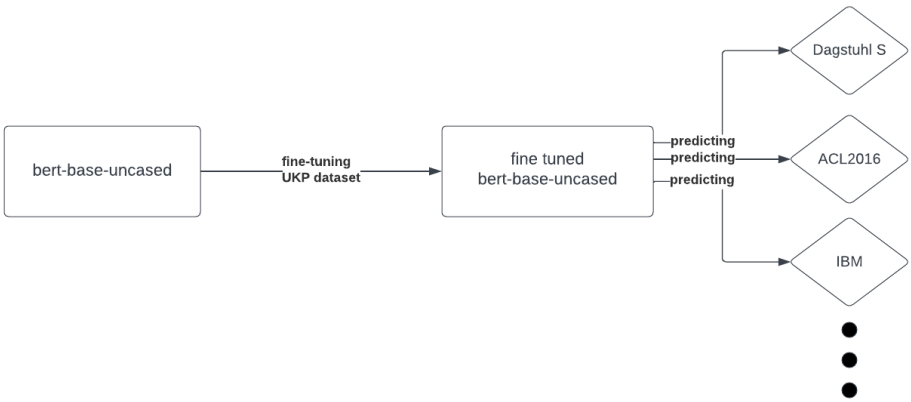

Figure 4.1: An overview of the experimental setup using bert-base-uncased as a base model and the UKP dataset as a source data set.

With these predictions at hand, one can compute the aforementioned metrics: accuracy or Cohens Kappa score.

4.1. TRAINING

To ensure appropriate training, a learning rate of 2e-5 was chosen to achieve better convergence, as commonly used with Bert models. Regarding the number of epochs, several metrics were monitored indicating that five epochs were sufficient to learn the classification in the sufficiency dimensions. Results also show that there were no major accuracy improvements past the fourth epoch in all experiments. All fine-tuned models are accessible through their respective hugging face pages[1]. Each data set was split into 80/20 train test sets using a fixed seed. If a dataset labeling was unbalanced by more than 10 percent (e.g. label 1 60%) oversampling of the smaller class was conducted to achieve balance. Table 4.1 shows the original label distributions. Furthermore, it was verified that no significant amount of samples was truncated for each data set.

[1] https://huggingface.co/jakub014

	Label 0	Label 1	Total	Label 0 (%)	Label 1 (%)
acl_pairs.csv	5872	5778	11650	50.40%	49.60%
acl_singles.csv	11650	11650	23300	50.00%	50.00%
dagstuhlE_singles.csv	194	121	315	61.59%	38.41%
dagstuhlE_pairs.csv	11737	11737	23474	50.00%	50.00%
dagstuhlS_singles.csv	188	127	315	59.68%	40.32%
dagstuhlS_Pairs.csv	11938	11938	23876	50.00%	50.00%
ibm_pairs.csv	3031	2666	5697	53.20%	46.80%
ibm_singles.csv	5697	5697	11394	50.00%	50.00%
redditCMV_pairs.csv	5151	5152	10303	50.00%	50.00%
redditCMV_singles.csv	10303	10303	20606	50.00%	50.00%
UKP_data.csv	348	681	1029	33.82%	66.18%
UKP_Pairs.csv	11250	11250	22500	50.00%	50.00%
redditCMVP_pairs	1728	1728	3456	50.00%	50.00%
redditCMVP_single	3456	3456	6912	50.00%	50.00%

Table 4.1: The label distributions for each data set used in the experiments.

As previously mentioned, several training curves were manually inspected to aim for appropriate training conditions. Figure 4.2 depicts one of such curves. The training loss decreases over time while the validation loss increases. Through that one can observe a trend toward over-fitting on the data. Since the model version with the best performance on the validation set is eventually selected this training picked the 'best' model at epoch 2. The tendency to over-fit on the training data is even more amplified for data sets with lower sample sizes. For instance, the dagstuhl data set consists of only 315 samples resulting in over-fitting in the early epochs already. Nevertheless, the training parameters were kept consistent due to showing promising metrics for most data sets when using bert-base-uncased as a base model.

Figure 4.2: The training curves for the IBM-argQ-30k base model fine-tuning on acl2016.

4.2. DATA ADJUSTMENTS

Since the data sets come in different formats adjustments were required to make them compatible with the aforementioned Task 1 or 2. These adjustments are explained in the

following paragraphs. After each paragraph one sample of the data set is provided. The appendix entails three stages: original, single task, and pair task format for each of the shown samples.

ACL2016 The data set depicts the dimension of convincingness. It has the format of two arguments with a label selecting one of the two. The data is split into multiple files, each being categorized by a corresponding discussion topic. The discussion topic was manually adjusted to a sentence. The mapping of topic to topic sentences can be found in the appendix. The topic sentence was then prepended to each argument keeping the original label. The data from all files was aggregated, matching the data format for Task 2 (pairs). To allow predictions from other models on this data set it also needs to be available for Task 1. This means changing the format to a single argument with a binary label. To achieve this the data was simplified by distributing the label 1 to arguments that were chosen out of the pair and giving the other argument the label 0. If an argument appeared in multiple pairs majority voting was conducted to decide its label. If the voting did not result in a decision the comment was excluded from the single format of this data set. The topic sentence is kept in this format.

> **A0** School uniform is a bad idea. In comparison to civil dress, school uniforms prove to be futile and wasteful once the child is out of school.
>
> **A1** School uniform is a bad idea. School uniforms are a BAD idea. Kids won't be able to show their color.
> **Label:** 0 (argument A0 is more convincing than A1)

IBM Debater ACL 2019 The data set is also annotated for convincingness. Similarly to the previous procedure, the data set is also provided in an argument pair format but already aggregated in one file. The topic sentence is already fitting such that prepending it to each argument and taking the label was a logical choice. Once again to transform the data set from pairs to singles we simplify by giving label 1 arguments that were selected and label 0 to the remaining one. To solve the problem of arguments occurring in several pairs we still follow the procedure described for the IBM data set. The topic sentence is once again kept in this format.

> **A0** We should ban human cloning. Following the announcement, then-White House Press Secretary Scott McClellan spoke on behalf of president George W. Bush and said that human cloning was "deeply troubling" to most Americans.
>
> **A1** We should ban human cloning. The U.N. General Assembly then voted on a non-binding resolution, calling upon all nations to "prohibit all forms of human cloning inasmuch as they are incompatible with human dignity and the protection of human life".
> **Label:** 1 (argument A1 is more convincing than A0)

Dagstuhl 15512 Argquality corpus The corpus devised by Wachsmuth et al. It contains annotations for several argument quality dimensions which we use for sufficiency and effectiveness. The argument is extracted together with a quality dimension which is named identically to the relevant field. The labeling is ternary in this data set. Every dimension is labeled from 1 to 3, from low to high. To adjust to other data sets the labeling needed to be simplified where for both effectiveness and sufficiency were changed to where *medium* is effective or sufficient. The change this data set to pairs the set of all arguments with label 1 and the set of all arguments having label 0 were used to perform a cartesian product. This results in pairs of arguments of the required format.

> **A0** Bottled water is somewhat less likely to be found in developing countries, where public water is least safe to drink. Many government programs regularly disperse bottled water for various reasons. Distributing small bottles of water is much easier than distributing large bulk storages of water. Also contamination from large water storage containers is much more likely than from single 12-20 ounce bottles of water.
> **Label sufficiency:** 1 (argument A0 is sufficient)
> **Label effectiveness:** 1 (argument A0 is effective)

RedditCMV Effectiveness The data set is provided in a format similar to a forum post. There exists a main thread that you can comment on (including chains of comments). The fields relevant fields extracted are *submission/title* for the title, *nodelta_comment/comments/0/body* for the *ineffective* comment as well as *delta_comment/comments/0/body* for the effective comment. Once again the title was prepended to each argument resulting in the appropriate pair format. To obtain the single argument format we once again simplify by deeming one comment effective and the other ineffective.

> **A0** I don't believe that parents should be able to name children whatever they want. CMV.. There are many factors in play when someone is bullied and it isn't just because you have an odd first name. If your argument is to prevent bullying then you'll have to include altering last names, ""foreign"" names, clothing, glasses, height, weight, IQ and financial status. The parents will also have to follow guidelines to make sure they don't drink to excess don't divorce or have affairs and also have ""approved"" jobs and homes. If one is unhappy with the name they have been given it can be legally changed later in life is one believes it will impede the chance of getting a good job. But parents can choose whatever name they wish for their child. I think it would serve society better to change the behavior of the bully not the name of the victim.

> **A1** I don't believe that parents should be able to name children whatever they want. CMV.. Freakonomics did a podcast, titled ""How Much Does Your Name Matter?"", on this very subject. The overall conclusion is that what your parents name you has virtually no effect on the outcome of your life. You might think ""La-a"" is a cruel name to give a kid, but that girl is will

statistically have the same chance of success and happiness in life as a girl named Whitney or Ashley with the same socioeconomic background.
Label: 1 (argument A1 is more effective than A0)

RedditCMV Persuasiveness Similarly to the previous data set, there exists a main thread that you can comment on (including chains of comments) but for this data set the data was already provided in pairs of argument chains. One being successful, the other one not. The data was furthermore simplified by only taking the first argument in each chain. Once again the title was prepended to each argument resulting in the appropriate pair format. To obtain the single argument format we once again simplify by deeming one comment effective and the other ineffective.

A0 CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone. It can signal condescension towards the person you are replying to, but it can also be a part of a joke where it is directed towards something or someone else your comment is talking about. Occasionally it is neither, but an understandable emphasis or intonation. It should be pretty easy to figure out based on the context.

A1 CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone. I just checked my post history, and the last time I used italics was to stress a point, and I think it works quite well for this purpose. It was in /r/askphilosophy and some kid wanted help with a paper. He said that he would cite anyone who helped him in his paper. I honestly don't have words for how bad an idea it is to cite random anonymous strangers from the internet in a paper. Putting the never in italics put extra emphasis on that this not something you should ever do. I hope I got the point across without having to call him an idiot or be condescending, but just telling him that it is a really, *really* bad idea."
Label: 0 (argument A0 is more persuasive than A1)

UKP The UKP data set revolves around sufficiency and is already present in a fitting single argument format with binary labeling. The fields *TEXT* and *ANNOTATION* were extracted. Where the labeling of *insufficient* and *sufficient* was mapped to 0 and 1. To obtain the format of pairs for this data set once again a cartesian product was calculated but since this data set is larger the amount of positive as well as negative samples were limited to 150 for each set, resulting in 22500 pairs.

A0 First and foremost , email can be count as one of the most beneficial results of modern technology . Many years ago , peoples had to pay a great deal of money to post their letters , and their payements were related to the weight of their letters or boxes , and many accidents may cause problem that the post could not be delivered. But nowadays , all people can take advantage of internet to have their own email free , and send their emails to everyone in no time , besides they can be sure if their emails have been delivered or not.

Label: 1 (argument A0 is sufficient)

To ensure the stability of models the fine-tuned models from bert-base-uncased were trained and tested on five different seeds. The model's accuracies from testing can be found in the appendix. No large disparities to the first found accuracy were found determining the models to be stable enough. Hyper-parameters were not tuned beyond adjustments through manual observations.

5

RESULTS

5.1. PERFORMANCE ACROSS AND WITHIN DIMENSIONS

The results obtained by using bert-base-uncased as a base model are shown in the following, see Table 5.1 for raw data. The columns depict the data set the model was fine-tuned on while the rows show the data set that was predicted. When inspecting the prediction accuracies it is possible to see that no model performs better at predicting data **within** its dimension, that it was not trained on, than **across** dimensions. Another finding worth mentioning is that the model trained on Dagstuhl's sufficiency dimension surprisingly scores the lowest accuracy (36,8%) of all predictions even though it is a prediction on the other sufficiency data set (**within** dimension). The converse is also performing poorly achieving only 42,5% accuracy the second lowest score. Sufficiency performs the worst out of all when it comes to predictions **within** the same dimension. The model trained on the ACL2016 data set achieved the highest average accuracy across all models with 62.3%. Most predictions are in the range of 50% to 60% with the exception of the performance on the data set it was trained on, 92%, and surprisingly the Dagstuhl sufficiency data, 74.2%. The model trained on ACL2016 was found to be the best at predicting the dagstuhl sufficiency and ACL2016 data whereas for most other data sets the corresponding highest-scoring model was the one that trained on the data itself. These findings already hint at the possibility of other aspects than the depicted quality dimension being more relevant to be able to generalize from one data set to another.

Another interesting finding is that the Dagstuhl sufficiency model scores best at predicting the Dagstuhl effectiveness data set (67% accuracy). This is most likely due to the correlation of the data itself. Wachsmuth et al. already have shown that the effectiveness annotations with the sufficiency annotations have a correlation coefficient of 0.73. The models trained on Dagstuhl data were the only ones that achieved prediction accuracies, on other data sets, that were greater than the accuracy on the data set the model was trained on.

To further compare the generalizability of data sets **within** dimensions and data sets **across** dimensions, we define the term base accuracy M as the percentage of correct

		Source dataset (trained on)							
		Sufficiency		Convincingness		Effectiveness		Persuasiveness	Average
		UKP	Dagstuhl S	ACL2016	IBM	RedditCMV	Dagstuhl E	RedditCMV_P	
Target dataset (tested on)	UKP	0.888	0.368	0.557	0.537	0.495	0.487	0.495	0.547
	Dagstuhl S	0.425	0.603	0.742	0.547	0.499	0.737	0.499	0.579
	ACL2016	0.488	0.523	0.920	0.579	0.552	0.541	0.552	0.594
	IBM	0.566	0.500	0.595	0.751	0.463	0.500	0.463	0.548
	RedditCMV	0.487	0.499	0.506	0.506	0.637	0.488	0.520	0.520
	Dagstuhl E	0.444	0.670	0.539	0.539	0.498	0.619	0.498	0.544
	RedditCMV_P	0.487	0.499	0.506	0.506	0.552	0.488	0.620	0.522
	Average	0.541	0.523	0.623	0.566	0.528	0.551	0.521	0.564

Table 5.1: The prediction accuracies across data sets using fine-tuned bert base uncased.

predictions on the data a model was trained on. For instance, the UKP data set was used to train a model, the model's accuracy on the test split of the UKP data is then considered the base accuracy M, 88.6%, of the UKP model (fine-tuning on bert-base-uncased). We also define N as the accuracy score on predicting another data set of the same dimension as the model's training data. As an example, the UKP model predicting the Dagstuhl sufficiency data would result in the accuracy N (42.5%). Finally, we consider the set K of accuracy scores of data sets that are not depicting the dimension of the data the model was trained on. For each model, one can look into the difference of accuracies $N-M$ and $K-M$ depicting the 'gain' in accuracy **within** and **across** dimensions. We do this to compare the gain or loss when doing predictions on other dimensions between models. This shows that on average the models have a 24.4% lower accuracy when predicting **within** dimensions on other data sets than predicting on the data the model was trained on ($N-M$). It also shows that on average models have a 19.2% lower accuracy when predicting **across** dimensions than predicting on the data the model was trained on ($K-M$). This is contrary to expectations and further implies that generalizations between data sets are complex.

Figures 5.1-5.3 depict the aforementioned differences in accuracies where a boxplot is used to show the differences **across** dimensions ($K-M$) and a star to depict the value $N-M$ for each data set. Indicated by the coloring as well as the background sections the diagrams are separated for each quality dimension. We can see that Dagstuhl S (sufficiency), Dagstuhl E (effectiveness), and ACL2016 (convincingness) all have outliers with, relatively, high values when predicting **across** dimensions resulting in more accuracy than predicting on its own data. We explain the high variation of results for the dagstuhl data set through the low sample size of 315. Generally, we would expect the stars (prediction **within** dimension) to be higher than the box plot (predictions **across** dimensions) for each data set since generalization **within** a quality dimension should be easier. We can see that this is only the case for the IBM data set in the dimension of convincingness. The result for the ACL2016 data set also shows that the prediction **within** dimensions is above the boxplot but not above its top outlier. Possibly showing a trend for the convincingness dimension. For the remaining data sets the difference N - M was either within or below the area of the box plot indicating a worse generalization for within dimensions.

Another aspect that was investigated was the agreement between predictions. If models were predicting only a fraction of the labels correctly it would be interesting if there is some overlap between them, if they have some form of agreement. For that the

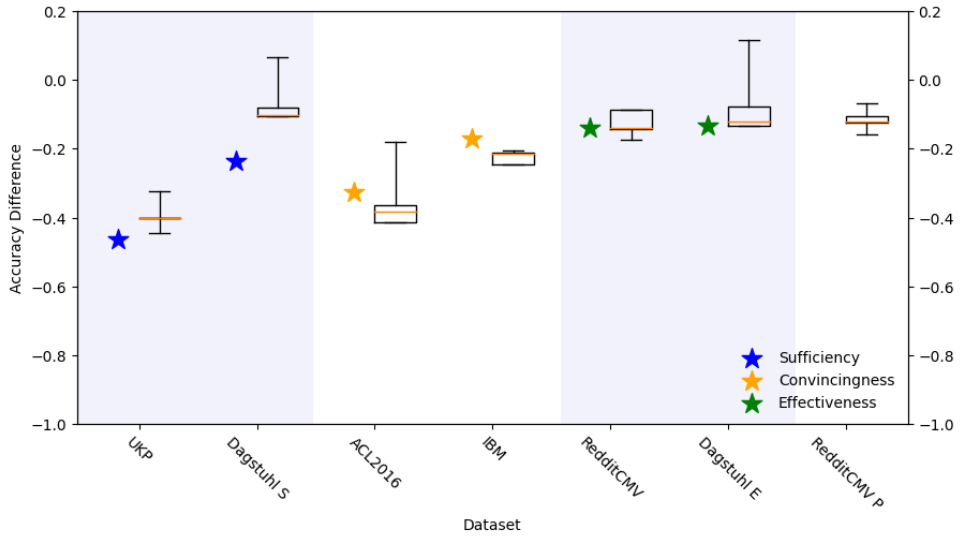

Figure 5.1: The difference of accuracies explained in the result section (bert-base-uncased).

Cohen's Kappa score [22] was used.

All predictions were collected and predictions on the same data set were evaluated with the Cohen's Kappa. Table 5.2 shows the total averaged scores for each combination of models using the base model bert-base-uncased. There are two separate tables since a model is only capable of predicting one of the two task types. One can see that the models DagstuhlE and DagstuhlS seem to agree to a fair degree which is quite logical since they were trained on the same text and labeling that has a high positive correlation. Interestingly we can observe that UKP has little to no agreement with any of the Dagstuhl models. Regarding the remaining models there seems to be little to no agreement between any of them, even though two of the data sets are from the same source, redditCMV. Nevertheless, it is no surprise since the accuracy from one to the other reddit

Average Cohens Kappa

	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.35	0.00
DagstuhlS		1.00	0.00
UKP			1.00

Average Cohens Kappa

	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	1.00	0.05	0.05	0.03
IBM		1.00	0.07	0.07
redditCMVE			1.00	0.09
redditCMVP				1.00

Table 5.2: The averaged Cohen's kappa score for each pair of models (with the same task and base model bert-base-uncased) across all predictions.

		Source dataset (trained on)							
		Sufficiency		Convincingness		Effectiveness		Persuasiveness	Average
		UKP	Dagstuhl S	ACL2016	IBM	RedditCMV	Dagstuhl E	RedditCMV_P	
Target dataset (tested on)	UKP	0.879	0.537	0.525	0.646	0.521	0.605	0.516	0.604
	Dagstuhl S	0.444	0.635	0.776	0.571	0.509	0.714	0.501	0.593
	ACL2016	0.506	0.602	0.930	0.577	0.594	0.574	0.561	0.621
	IBM	0.592	0.519	0.583	0.742	0.499	0.545	0.473	0.565
	RedditCMV	0.487	0.510	0.501	0.501	0.648	0.506	0.529	0.526
	Dagstuhl E	0.470	0.813	0.565	0.565	0.482	0.651	0.515	0.580
	RedditCMV_P	0.482	0.471	0.503	0.499	0.546	0.490	0.621	0.516
	Average	0.551	0.584	0.626	0.586	0.543	0.584	0.531	0.592

Table 5.3: The prediction accuracies across data sets using fine-tuned Cold-Fusion bert base.

Data set was not very good either.

5.2. IMPACT OF OTHER BASE MODELS

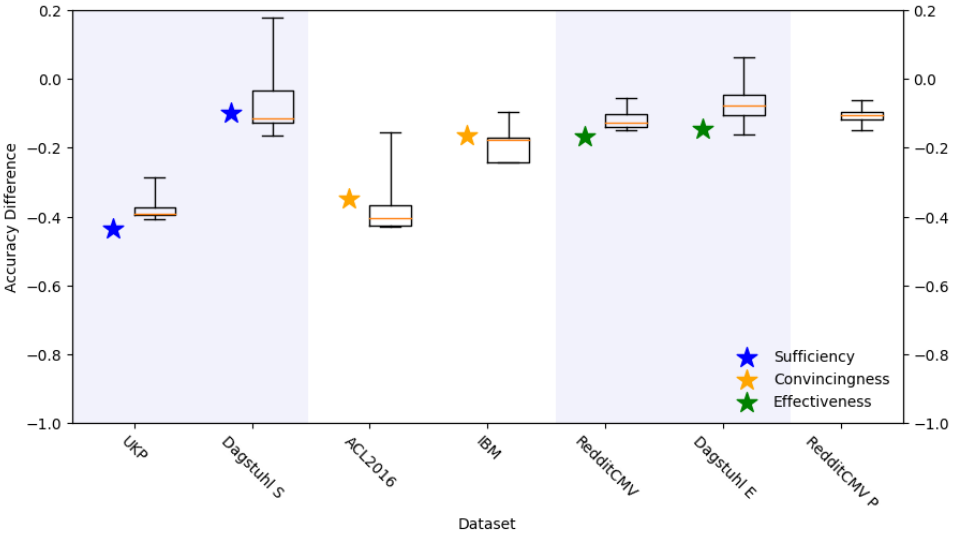

Figure 5.2: The difference of accuracies explained in the result section (Cold-Fusion).

The results obtained by using the ColD-Fusion and IBM ArgQ30k models as base models are shown in the following, see Tables 5.3 and 5.5 for raw data. When zooming out and comparing the performance of the aforementioned models to the first set of experiments one can find an overall increase in accuracy across all data sets of 3% when using Cold-fusion-bert-base-uncased and even over 4% when using the IBM ArgQ30k-bert-base-uncased model. This speaks towards a gain when the source and target data sets are well aligned. Especially when we consider the dimensions of argument quality being directly correlated to argument quality itself. It is also worth noting that the Dagstuhl sufficiency model performs significantly better with the intermediate IBM ArgQ30k model, obtaining an almost 9% better accuracy overall. It gains the most improvement

Average Cohens Kappa

	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.41	0.15
DagstuhlS		1.00	0.03
UKP			1.00

Average Cohens Kappa

	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	1.00	0.27	0.11	0.05
IBM		1.00	0.02	0.08
redditCMVE			1.00	0.11
redditCMVP				1.00

Table 5.4: The averaged Cohen's kappa score for each pair of models (with the same task and base model ColdFusion bert base) across all predictions.

		Source dataset (trained on)							
		Sufficiency		Convincingness		Effectiveness		Persuasiveness	Average
		UKP	Dagstuhl S	ACL2016	IBM	RedditCMV	Dagstuhl E	RedditCMV_P	
Target dataset (tested on)	UKP	0.884	0.650	0.518	0.616	0.514	0.646	0.512	0.620
	Dagstuhl S	0.498	0.698	0.733	0.633	0.552	0.759	0.498	0.625
	ACL2016	0.577	0.622	0.927	0.705	0.612	0.613	0.561	0.659
	IBM	0.550	0.534	0.641	0.760	0.546	0.538	0.493	0.580
	RedditCMV	0.495	0.518	0.533	0.533	0.653	0.517	0.523	0.539
	Dagstuhl E	0.505	0.743	0.622	0.622	0.517	0.730	0.497	0.605
	RedditCMV_P	0.489	0.515	0.513	0.516	0.554	0.492	0.624	0.529
	Average	0.571	0.612	0.641	0.626	0.564	0.613	0.530	0.620

Table 5.5: The prediction accuracies across data sets using fine-tuned IBM ArgQ30k.

Average Cohens Kappa

	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.72	0.23
DagstuhlS		1.00	0.23
UKP			1.00

Average Cohens Kappa

	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	1.00	0.39	0.23	0.03
IBM		1.00	0.29	0.09
redditCMVE			1.00	0.08
redditCMVP				1.00

Table 5.6: The averaged Cohen's kappa score for each pair of models (with the same task and base model IBM ArgQ30k) across all predictions.

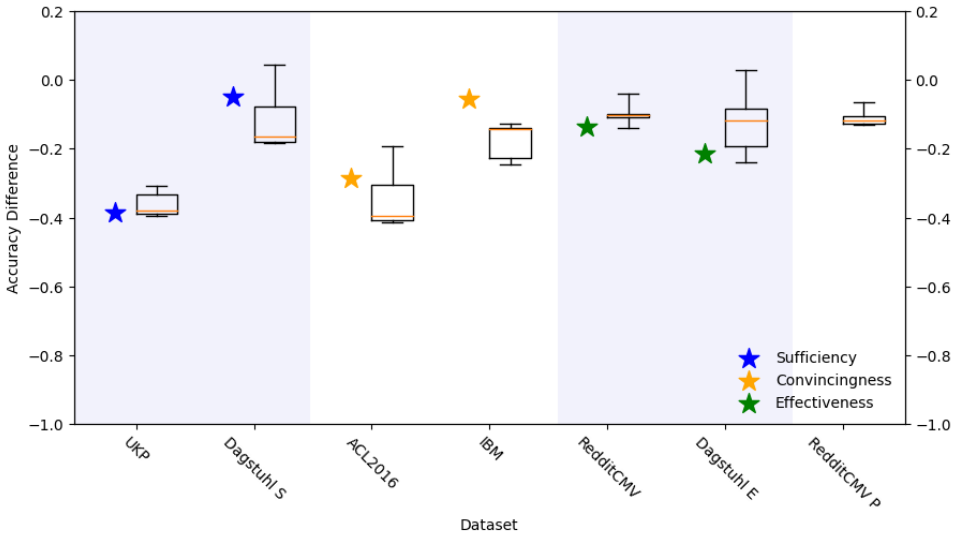

Figure 5.3: The difference of accuracies explained in the result section (IBM ArgQ30k).

at predicting within the sufficiency dimension (+28,2%). Once again looking into the difference in accuracies of predictions on the data set the model was trained on and either the data sets **within** or **across** dimensions, averaged if there are multiple. We can see a trend similar to the one found before in the results of the ColD-Fusion model. An average difference, or loss, in the accuracy of 22.6% when predicting **within** dimensions and 17.9% when predicting **across** dimensions. When looking into IBM ArgQ30k on the other hand we can observe a loss in accuracy of roughly 18.7% for both **within** and **across** dimensions. While still not being in line with previous assumptions the finding for IBM ArgQ30k indicates a potential indifference between predictions **within** and **across** dimensions.

We can once again look at the difference in accuracies in the box plot diagrams in Figures 5.1-5.3. It is possible to see that the results with Cold-Fusion as a base model (Figure 5.2) show similar trends to the results with Bert-base as a base model. One thing to note is that the predictions within the dimension of sufficiency seem to lose slightly less accuracy. As for the box plots from the IBM ArgQ30k results one can see even more amplified trends. The sufficiency predictions **within** dimensions have improved to the point where the DagstuhlS prediction **within** dimensions is above the box (Figure 5.3, blue star). One can once again see a clearer separation between **within** and **across** for the convincingness dimension. As for the dimension of effectiveness, one can observe a slight trend of losing accuracy. The box plots in this diagram are also shorter due to a smaller variance in the accuracies **across** dimensions.

Overall one can see that the box plots across all three diagrams stay around the same accuracy difference (height). This suggests that the training performance M and predictions on other dimensions K have a positive correlation.

As for the agreement of models for these base models (Tables 5.4 and 5.6) we can

see a trend that agreement increases which makes sense since overall accuracy also increased. Nevertheless one can see larger improvements up to a 0.72 agreement score between the Dagstuhl models with the IBM ArgQ30k base model. According to the traditional interpretation of the metric this means there exists substantial agreement between the models. IBM and ACL2016 also reached some agreement with these base models which can be interpreted as fair to moderate. The RedditCMVE model also obtained a slight to fair agreement with the ACL2016 and IBM models. Overall one can argue whether a form of generalization can be observed but it is clear that there exists at least some overlap in the understanding that some models have gained through training.

To further strengthen the displayed findings regarding RQ1.1 we conducted a two sided t-test for each base model with the null hypothesis that there is no difference between the results **within** dimensions and the results **across** dimensions. For that, we gathered the accuracy scores when predicting data sets **within** the model's dimension and **across** the model's dimension. We use a t-test since we assume that with sufficiently many data sets and predictions the results should approach a normal distribution. We obtain the p-values 0.51, 0.78, 0.41 respectively when using the results from the base models: bert-base-uncased, ColD-Fusion and IBM ArgQ30k. We can not find any strong evidence against the null hypothesis such that we accept it.

6

Discussion and Conclusion

This research aimed to investigate the generalizability of argument quality dimensions. Our experimental design incorporated the use of NLP models and intermediate training. The data collected from the models' predictions were carefully analyzed using their accuracy. Furthermore, the agreement between models, investigating the relation between dimensions of argument quality, was analyzed using the Cohens Kappa metric.

The findings of this study reveal a notable difference from expectations. Regarding RQ1.1, the data consistently demonstrate that predictions within dimensions were not better than across dimensions. There have been little to no findings on the generalizability of argument quality dimensions through the trained models. As for RQ1.2, using different base models still resulted in quite similar trends. These results suggest that beyond the theoretical framework of quality dimensions by Wachsmuth et al. there are more important factors relevant to predicting quality dimensions using language models.

It is crucial to recognize the limitations of this study. Possible sources of error, alternative explanations, or confounding factors need to be considered. Starting with the language models themselves. There are a plethora of parameters one can explore: different architectures, more optimized training, or hyperparameter optimization. Beyond that, one can also think of different intermediate training, for instance, only training on general argument quality data. The quality dimensions were also not exhausted in this study, for once the selected ones were simplified to fit a binary classification task. One could study how to measure an individual dimension on its own and would most likely still not end up with a definitive answer. Secondly, there are several other dimensions that were untouched in this study leaving a large space to still explore. Regarding data used in this study, there were several decisions made compromising the data quality and lowering its depiction of the respective quality dimension. For instance when converting a data set from pairs to single arguments. A very naive approach was taken to convert these pairs of arguments. Oftentimes it is not correct to assume that if one argument is, for instance, less convincing that it is not convincing at all. Beyond that, we can also say that some data sets still had relevant differences. For instance, some data sets had fields

about the argument's topic, title, or even premise while other arguments were provided without context. Even the data sets with given premises had differences in the way they were formulated. Some topics were phrased such that the argument was in favor of it while other titles described views that should be changed through the made arguments. Some of these titles were questions while others were statements or even just a topic. The arguments themselves also fundamentally varied in their language and style, due to stemming from different domains ranging from online forums to student essays. Arguments also varied a lot in their length even within some of the data sets, like the Reddit ones making them especially difficult to learn. Staying with the complexity of some of the data sets, while manually inspecting some of the samples I also realized that grasping why one argument should be more effective than the others was far from trivial. For instance, for the domain of sufficiency, one would need to assess very carefully what premise is argued for and why it is less or more sufficient than another argument. This all just shows how much to explore there still is.

Future research should aim to replicate such findings using different methodologies or explore additional variables that could provide further insights into the observed discrepancy between theory and data. The models trained in this study are all available on their respective hugging face pages. Variables one might inspect could be varying sizes of data or a similar study but with a focus on what domain the arguments are from. Naturally, the aforementioned limitations are also paths for further exploration.

Moving a bit further, assuming that argument quality dimension estimations would be quite accurate one can then look into the relation with other tasks. For example, one could investigate the task of identifying whether an argument is for or against a claim. Being able to already estimate argument quality dimensions could provide valuable insights for that task.

Taking it even further, by being able to assess argument quality dimensions one should be able to understand arguments in general better. This would open up directions towards enhancing or explaining arguments but also findings gaps in argumentation and possibly generating counterarguments. Investigating different domains of arguments could also be interesting then. Maybe one can show that certain domains are more likely to persuade people through their dialectic abilities rather than logic or the other way around. However, for most of these ideas, there exists one central problem, namely the lack of sufficient quality data. It would either be costly or even unobtainable. Fortunately, there also exists the idea of having a human in the loop to mitigate or even avoid such problems. Being able to steer the model in the right direction, having the benefits of human judgment while also performing tasks efficiently could play a significant role in shaping how technology is developed.

To conclude this study, the implications of the findings are relevant, opening avenues for further investigation. In conclusion, this research has presented results that do not show generalizations of argument quality dimensions through NLP models.

ACKNOWLEDGEMENTS

I am very grateful for the extraordinary guidance and support provided to me throughout the course of this thesis. Without the insightful contributions and guidance of my supervisors, this achievement would not have been possible.

First and foremost, I would like to express my gratitude to Michiel, whose dedication to mentorship has been instrumental in shaping my understanding and approach. Weekly encouragement, thoughtful guidance, and expertise have been a source of motivation and inspiration for me. Your ability to convey feedback and provide clear perspectives has truly enriched my journey.

I am also indebted to Pradeep and Catholijn, whose attention to detail has profoundly influenced my work. Your thoughtful critique, constructive feedback, and willingness to engage in in-depth discussions have been invaluable to me. Thank you for giving me the opportunity of conducting my thesis in this field.

Last but not least, I express my gratitude to my family and friends for their unending encouragement and patience throughout this journey. The lessons learned and experiences gained will certainly positively impact my future endeavors.

Gratefully,
Jakub Nguyen

BIBLIOGRAPHY

[1] J. S. Fishkin, *The Voice of the People: Public Opinion and Democracy*, en. Yale University Press, 1995, ISBN: 978-0-300-06556-5.

[2] S. R. Arnstein, "A ladder of citizen participation," *Journal of the American Institute of Planners*, vol. 35, no. 4, pp. 216–224, 1969. DOI: 10.1080/01944366908977225. eprint: https://doi.org/10.1080/01944366908977225. [Online]. Available: https://doi.org/10.1080/01944366908977225.

[3] C. Stab and I. Gurevych, "Identifying Argumentative Discourse Structures in Persuasive Essays," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 46–56. DOI: 10.3115/v1/D14-1006. [Online]. Available: https://aclanthology.org/D14-1006 (visited on 01/26/2023).

[4] H. Wachsmuth, B. Stein, and Y. Ajjour, ""PageRank" for Argument Relevance," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1117–1127. [Online]. Available: https://aclanthology.org/E17-1105 (visited on 01/26/2023).

[5] R. El Baff, H. Wachsmuth, K. Al-Khatib, and B. Stein, "Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 454–464. DOI: 10.18653/v1/K18-1044. [Online]. Available: https://aclanthology.org/K18-1044 (visited on 01/26/2023).

[6] I. Persing, A. Davis, and V. Ng, "Modeling Organization in Student Essays," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 229–239. [Online]. Available: https://aclanthology.org/D10-1023 (visited on 01/26/2023).

[7] H. Wachsmuth, K. Al-Khatib, and B. Stein, "Using Argument Mining to Assess the Argumentation Quality of Essays," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1680–1691. [Online]. Available: https://aclanthology.org/C16-1158 (visited on 01/26/2023).

[8] M. Potthast, L. Gienapp, F. Euchner, *et al.*, "Argument Search: Assessing Argument Relevance," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19, New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 1117–1120, ISBN: 978-1-4503-6172-9. DOI: 10.1145/3331184.3331327. [Online]. Available: https://doi.org/10.1145/3331184.3331327 (visited on 01/25/2023).

[9] H. Wachsmuth, N. Naderi, Y. Hou, *et al.*, "Computational Argumentation Quality Assessment in Natural Language," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187. [Online]. Available: https://aclanthology.org/E17-1017 (visited on 01/26/2023).

[10] W. Carlile, N. Gurrapadi, Z. Ke, and V. Ng, "Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 621–631. DOI: 10.18653/v1/P18-1058. [Online]. Available: https://aclanthology.org/P18-1058 (visited on 04/25/2023).

[11] A. Bondarenko, M. Fröbe, J. Kiesel, *et al.*, "Overview of Touché 2022: Argument Retrieval," en, in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2022, pp. 311–336, ISBN: 978-3-031-13643-6. DOI: 10.1007/978-3-031-13643-6_21.

[12] S. Gretz, R. Friedman, E. Cohen-Karlik, *et al.*, *A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis*, arXiv:1911.11408 [cs] version: 1, Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.11408 (visited on 04/26/2023).

[13] Z. Rahimi, D. J. Litman, R. Correnti, L. C. Matsumura, E. Wang, and Z. Kisa, "Automatic Scoring of an Analytical Response-To-Text Assessment," en, in *Intelligent Tutoring Systems*, D. Hutchison, T. Kanade, J. Kittler, *et al.*, Eds., vol. 8474, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 601–610, ISBN: 978-3-319-07220-3 978-3-319-07221-0. DOI: 10.1007/978-3-319-07221-0_76. [Online]. Available: http://link.springer.com/10.1007/978-3-319-07221-0_76 (visited on 02/06/2023).

[14] C. Stab and I. Gurevych, "Recognizing Insufficiently Supported Arguments in Argumentative Essays," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 980–990. [Online]. Available: https://aclanthology.org/E17-1092 (visited on 01/26/2023).

[15] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee, "Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Republic and Canton of Geneva, CHE: International World Wide Web Confer-

ences Steering Committee, Apr. 2016, pp. 613–624, ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883081. [Online]. Available: https://doi.org/10.1145/2872427.2883081 (visited on 02/05/2023).

[16] A. Lauscher, L. Ng, C. Napoles, and J. Tetreault, "Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4563–4574. DOI: 10.18653/v1/2020.coling-main.402. [Online]. Available: https://aclanthology.org/2020.coling-main.402 (visited on 07/03/2023).

[17] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18. DOI: 10.18653/v1/N19-5004. [Online]. Available: https://aclanthology.org/N19-5004.

[18] A. Malte and P. Ratadiya, *Evolution of transfer learning in natural language processing*, Number: arXiv:1910.07370 arXiv:1910.07370 [cs], Oct. 2019. DOI: 10.48550/arXiv.1910.07370. [Online]. Available: http://arxiv.org/abs/1910.07370 (visited on 08/02/2023).

[19] University of St.Gallen (HSG), Institute of Information Management, St.Gallen, Switzerland, T. Wambsganss, N. Molyndris, M. Söllner, and University of Kassel, Information Systems and Systems Engineering, Kassel, Germany, "Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach," en, in *WI2020 Zentrale Tracks*, GITO Verlag, Mar. 2020, pp. 341–356, ISBN: 978-3-95545-335-0. DOI: 10.30844/wi_2020_c9-wambsganss. [Online]. Available: https://library.gito.de/2021/07/13/wi2020-zentrale-tracks-24/ (visited on 08/02/2023).

[20] X. Hua and L. Wang, "Efficient Argument Structure Extraction with Transfer Learning and Active Learning," in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 423–437. DOI: 10.18653/v1/2022.findings-acl.36. [Online]. Available: https://aclanthology.org/2022.findings-acl.36 (visited on 08/02/2023).

[21] L. Choshen, E. Venezian, S. Don-Yehia, N. Slonim, and Y. Katz, *Where to start? Analyzing the potential value of intermediate models*, arXiv:2211.00107 [cs], Nov. 2022. [Online]. Available: http://arxiv.org/abs/2211.00107 (visited on 05/01/2023).

[22] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *International Conference on Fuzzy Systems*, 2010, pp. 1–8. DOI: 10.1109/FUZZY.2010.5584447.

Seed exploration							
	Sufficiency		Convincingness		Effectiveness		Persuasiveness
Run	UKP	Dagstuhl S	ACL2016	IBM	RedditCMV	Dagstuhl E	RedditCMV_P
1	0,888	0,603	0,920	0,751	0,637	0,619	0,620
2	0,923	0,615	0,912	0,694	0,619	0,632	0,627
3	0,907	0,627	0,904	0,723	0,605	0,613	0,609
4	0,896	0,630	0,896	0,695	0,588	0,604	0,608
5	0,861	0,636	0,888	0,712	0,572	0,593	0,602

Figure 1: The stability of the fine-tuning on bert-base-uncased was verified by fine-tuning using five different layer initializations.

Raw sample from ACL2016

#id: arg199550_arg554188
label: a1,
a1: In comparison to civil dress, school uniforms prove to be futile and wasteful once the child is out of school.
a2: School uniforms are a BAD idea. Kids won't be able to show their color.

Sample from ACL2016 adjusted for single task

a1: School uniform is a bad idea. In comparison to civil dress, school uniforms prove to be futile and wasteful once the child is out of school.
label: 1 (convincing)

Sample from ACL2016 adjusted for pair task

a0: School uniform is a bad idea. In comparison to civil dress, school uniforms prove to be futile and wasteful once the child is out of school.
a1: School uniform is a bad idea. School uniforms are a BAD idea. Kids won't be able to show their color. Label: 0
label: 0 (a0 is more convincing)

Raw sample from IBM Debater

topic: We should ban human cloning,
a0: Following the announcement, then-White House Press Secretary Scott McClellan spoke on behalf of president George W. Bush and said that human cloning was "deeply troubling" to most Americans.
a1: The U.N. General Assembly then voted on a nonbinding resolution, calling upon all nations to prohibit all forms of human cloning inasmuch as they are incompatible with human dignity and the protection of human life.
label: 1
acceptance_rate: 1.0
a0_stance: PRO
a1_stance: PRO
a0_detection_score: 0.8052822351455688
a1_detection_score: 0.6252058148384094

a0_id: 5360
a1_id: 5319
a0_wikipedia_article_name: Ethics of cloning,
a1_wikipedia_article_name: Genetics Policy Institute,
a0_wikipedia_url: https://en.wikipedia.org/wiki/Ethics_of_cloning,
a1_wikipedia_url: https://en.wikipedia.org/wiki/Genetics_Policy_Institute

Sample from IBM Debater adjusted for single task

a0: We should ban human cloning. Following the announcement, then-White House Press Secretary Scott McClellan spoke on behalf of president George W. Bush and said that human cloning was "deeply troubling" to most Americans.
label: 0 (not convincing)

Sample from IBM Debater adjusted for pair task

a0: We should ban human cloning. Following the announcement, then-White House Press Secretary Scott McClellan spoke on behalf of president George W. Bush and said that human cloning was "deeply troubling" to most Americans.
a1: We should ban human cloning. The U.N. General Assembly then voted on a nonbinding resolution, calling upon all nations to "prohibit all forms of human cloning inasmuch as they are incompatible with human dignity and the protection of human life."
label: 1 (a1 more convincing)

Raw sample from Dagstuhl

annotator 1
argumentative y
overall quality 1 (Low)
local acceptability 1 (Low)
appropriateness 1 (Low)
arrangement 1 (Low)
clarity 2 (Average)
cogency 1 (Low)
effectiveness 1 (Low)
global acceptability 1 (Low)
global relevance 1 (Low)
global sufficiency 1 (Low)
reasonableness 1 (Low)
local relevance 1 (Low)
credibility 1 (Low)
emotional appeal 1 (Low)
sufficiency 1 (Low)
argument: it is true that bottled water is a waste, but bottles can be reused!

#id arg219250
issue ban-plastic-water-bottles
stance no-bad-for-the-economy

Sample from Dagstuhl adjusted for single task

a0: it is true that bottled water is a waste, but bottles can be reused!
label effectiveness: 0
label sufficiency: 0

Sample from Dagstuhl adjusted for pair task

a0: it is true that bottled water is a waste, but bottles can be reused!
a1: Bottled water is somewhat less likely to be found in developing countries, where public water is least safe to drink. Many government programs regularly disperse bottled water for various reasons. Distributing small bottles of water is much easier than distributing large bulk storages of water. Also contamination from large water storage containers is much more likely than from single 12-20 ounce bottles of water.
label 1 (a1 is sufficient)

Raw sample from RedditCMV

"op_text":"I like to be brief, but it does specify 500 characters. Basically, to give an example:*Maybe* you're talking about the *preferred gender pronouns* because you *actually* hold a *sexist view*. I'm not quite sure why it irritates me so much, but whenever I see a comment like that, particularly when things like feminism or gender politics are discussed on here, it turns my stomach and makes me think the person writing it has swallowed a bunch of asterisks. I'm open to this view changing, it's quite a benign one. It just rubs me up the wrong way and *I'd love to change it*.

op_title:"CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone",

positive comment: It can signal condescension towards the person you are replying to, but it can also be a part of a joke where it is directed towards something or someone else your comment is talking about. Occasionally it is neither, but an understandable emphasis or intonation. It should be pretty easy to figure out based on the context.

negative comment: I just checked my post history, and the last time I used italics was to stress a point, and I think it works quite well for this purpose. It was in /r/askphilosophy and some kid wanted help with a paper. He said that he would cite anyone who helped him in his paper.I honestly don't have words for how bad an idea it is to cite random anonymous strangers from the internet in a paper. Putting the never in italics put extra emphasis on that this not something you should ever do. I hope I got the point across without having to call him an idiot or be condescending, but just telling him that it is a really, *really* bad idea.

Sample from RedditCMV adjusted for single task

a0: CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone. It can signal condescension towards the person you are replying to, but it can also be a part of a joke where it is directed towards something or someone else your comment is talking about. Occasionally it is neither, but an understandable emphasis or intonation. It should be pretty easy to figure out based on the context.
label: 1 (Persuasive)

Sample from RedditCMV adjusted for pair task

a0: CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone. It can signal condescension towards the person you are replying to, but it can also be a part of a joke where it is directed towards something or someone else your comment is talking about. Occasionally it is neither, but an understandable emphasis or intonation. It should be pretty easy to figure out based on the context.
a1: CMV: Using italics in an internet discussion forum signals condescension, and is a really rude way to reply to someone. I just checked my post history, and the last time I used italics was to stress a point, and I think it works quite well for this purpose. It was in /r/askphilosophy and some kid wanted help with a paper. He said that he would cite anyone who helped him in his paper. honestly don't have words for how bad an idea it is to cite random anonymous strangers from the internet in a paper. Putting the never in italics put extra emphasis on that this not something you should ever do. I hope I got the point across without having to call him an idiot or be condescending, but just telling him that it is a really, *really* bad idea. "
label: 0 (a0 more persuasive)

Raw sample from UKP

argument: First and foremost , email can be count as one of the most beneficial results of modern technology . Many years ago , peoples had to pay a great deal of money to post their letters , and their payements were related to the weight of their letters or boxes , and many accidents may cause problem that the post could not be delivered . But nowadays , all people can take advantage of internet to have their own email free , and send their emails to everyone in no time , besides they can be sure if their emails have been delivered or not.
label: sufficient

Sample from UKP adjusted for single task

a0: "First and foremost , email can be count as one of the most beneficial results of modern technology . Many years ago , peoples had to pay a great deal of money to post their letters , and their payements were related to the

weight of their letters or boxes , and many accidents may cause problem that the post could not be delivered . But nowadays , all people can take advantage of internet to have their own email free , and send their emails to everyone in no time , besides they can be sure if their emails have been delivered or not . "

label: 1 (sufficient)

Sample from UKP adjusted for pair task

a0: First and foremost , email can be count as one of the most beneficial results of modern technology . Many years ago , peoples had to pay a great deal of money to post their letters , and their payements were related to the weight of their letters or boxes , and many accidents may cause problem that the post could not be delivered . But nowadays , all people can take advantage of internet to have their own email free , and send their emails to everyone in no time , besides they can be sure if their emails have been delivered or not .

a1: Another important aspect on technology is transferring money . Today , students can apply for foreign universities much easier than before . Not only with the help of sending email , but also using credit cards to pay all necessary fees online . Therefore , with the advent of internet and online paying systems , you can do many thing at your home easily .

label: 0 (a0 is more sufficient)

Predictions on: acl_singles			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.37	-0.01
DagstuhlS	0.37	1.00	0.00
UKP	-0.01	0.00	1.00
Predictions on: DagstuhlE_test			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.27	0.07
DagstuhlS	0.27	1.00	0.02
UKP	0.07	0.02	1.00
Predictions on: DagstuhlS_test			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.27	0.07
DagstuhlS	0.27	1.00	0.02
UKP	0.07	0.02	1.00
Predictions on: ibm_singles			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	1.00	0.00
DagstuhlS	1.00	1.00	0.00
UKP	0.00	0.00	1.00
Predictions on: redditCMV_singles			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.15	-0.04
DagstuhlS	0.15	1.00	0.00
UKP	-0.04	0.00	1.00
Predictions on: UKP_test			
-	DagstuhlE	DagstuhlS	UKP
DagstuhlE	1.00	0.03	-0.08
DagstuhlS	0.03	1.00	-0.02
UKP	-0.08	-0.02	1.00

Figure 2: The Cohens kappa scores for each pair of models with the single argument task across all predictions.

Predictions on: ACL2016_test				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	15.91%	9.84%	8.43%
IBM	15.91%	100.00%	8.74%	9.21%
redditCMVE	9.84%	8.74%	100.00%	11.20%
redditCMVP	8.43%	9.21%	11.20%	100.00%
Predictions on: dagstuhlE_pairsMediumIsEffective				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	-2.31%	0.25%	-0.64%
IBM	-2.31%	100.00%	16.45%	13.16%
redditCMVE	0.25%	16.45%	100.00%	14.71%
redditCMVP	-0.64%	13.16%	14.71%	100.00%
Predictions on: dagstuhlS_MediumIsSufficientPairs				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	-2.25%	-0.04%	-0.52%
IBM	-2.25%	100.00%	14.90%	12.71%
redditCMVE	-0.04%	14.90%	100.00%	13.11%
redditCMVP	-0.52%	12.71%	13.11%	100.00%
Predictions on: IBM_debater_test				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	15.60%	7.02%	3.68%
IBM	15.60%	100.00%	-2.26%	1.16%
redditCMVE	7.02%	-2.26%	100.00%	3.16%
redditCMVP	3.68%	1.16%	3.16%	100.00%
Predictions on: RedditCMVE_test				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	3.46%	12.90%	9.28%
IBM	3.46%	100.00%	0.11%	1.71%
redditCMVE	12.90%	0.11%	100.00%	4.26%
redditCMVP	9.28%	1.71%	4.26%	100.00%
Predictions on: UKP_Pairs				
-	ACL2016	IBM	redditCMVE	redditCMVP
ACL2016	100.00%	-2.29%	0.52%	0.28%
IBM	-2.29%	100.00%	5.51%	4.78%
redditCMVE	0.52%	5.51%	100.00%	6.71%
redditCMVP	0.28%	4.78%	6.71%	100.00%

Figure 3: The Cohens kappa scores for each pair of models with the paired argument task across all predictions.