Reproducing the Argument Quality Prediction of Project Debater

Ines Zelch[1,2] **Matthias Hagen**[1] **Benno Stein**[3] **Johannes Kiesel**[4]

[1]Friedrich-Schiller-Universität Jena [2]Leipzig University [3]Bauhaus-Universität Weimar
[4]GESIS – Leibniz Institute for the Social Sciences

Abstract

A crucial task when analyzing arguments is to determine their quality. Especially when you have to choose from a large number of suitable arguments, the determination of a reliable argument quality value is of great benefit. Probably the best-known model for determining such an argument quality value was developed in IBM's Project Debater and made available to the research community free of charge via an API. In fact, the model was never open and the API is no longer available. In this paper, IBM's model is reproduced using the freely available training data and the description in the corresponding publication. Our reproduction achieves similar results on the test data as described in the original publication. Further, the predicted quality scores of reproduction and original show a very high correlation (Pearson's $r = 0.9$) on external data.

1 Introduction

When developing large datasets of arguments, the automatic assessment of the arguments' quality is crucial in order to provide arguments of sufficient quality for applications like a searchable argument index (Dumani and Schenkel, 2020; Wachsmuth et al., 2017b). A commonly used model for argument quality prediction was developed as part of the IBM argumentation system Project Debater (Bar-Haim et al., 2021; Slonim et al., 2021) and was made available for researchers via an API (Bar-Haim et al., 2021). The model was used, for example, by Bar-Haim et al. (2020) to select high quality arguments for the generation of key points, and by Alshomary and Wachsmuth (2023) for the generation of counter-arguments. However, as the API was closed in May 2024 (and the model is no longer available on request from the authors), this high-quality resource is no longer accessible to researchers; i.e., research based on this model cannot be applied to new datasets.

This paper contributes to an open reproduction of IBM's original model in order to make this important resource available again.[1] We follow the specifications of the original publication (Gretz et al., 2019) to finetune a BERT regression model on the publicly available original dataset of crowdsourced arguments and quality ratings.[2] As shown in this paper, our model achieves a very high correlation in terms of predicted quality scores with the IBM model: In a test with a subset of the third-party args.me corpus, the Pearson's r is 0.9.

The paper in hand outlines the retraining process and presents an analysis of the predictions of the trained models and a comparison with the original model. Section 2 provides an overview of the concept of argument quality in general and the IBM model of Gretz et al. (2019) in particular. Section 3 describes the reproduction of the IBM model in the detail. Section 4 reports on two evaluation studies on our reproduced model: (1) Using the original test data, we calculate the Pearson and Spearman correlation coefficients between the predictions of our model and the real annotations and compare these with the numbers given in the original paper (Gretz et al., 2019). (2) Using the args.me corpus (Ajjour et al., 2019), we calculate the same coefficients, but between the predictions of our model and the IBM model. For this purpose, we had acquired the necessary predictions of the IBM model before the API shut down. Interestingly, we find considerable differences in predictions for argumentative texts from the args.me corpus, although we achieve similar effectiveness with the original test set. This observation implies that the score achieved on a particular test set does not necessarily reflect the ability of the model to generalize to external data.

In order to keep the reproduced model lean and

[1] github.com/webis-de/argmining25-reproducing-ibm-arg-quality-api

[2] https://research.ibm.com/debating_data.shtml#Argument_Quality

Proceedings of the 12th Argument Mining Workshop, pages 181–188
July 31, 2025 ©2025 Association for Computational Linguistics

to make it usable in downstream applications without further dependencies on external models, we deliberately refrained from extending the IBM model by integrating LLMs.

2 Related Work

Argument quality can be assessed considering various quality dimensions. An overview of these dimensions is compiled in Wachsmuth et al. (2017a) and extended by Ivanova et al. (2024). They include logical dimensions that affect the cogency of an argument, dialectic dimensions that influence the reasonableness of arguments, and rhetoric dimensions that are important for an argument's effectiveness. Different quality dimensions are considered in existent datasets, mainly annotated in an absolute manner where each argument is labeled individually (Toledo et al., 2019; Ivanova et al., 2024). Other works approach argument quality analysis in a relative way, processing arguments in pairs and choosing the one of higher quality (Toledo et al., 2019). The latter approach has the advantage of being less complex (Ivanova et al., 2024), resulting in potentially more consistent annotations. Additionally, the various approaches applying absolute annotations often use different annotation scales that are not necessarily transferable to each other (Ivanova et al., 2024).

The argument quality model of the Project Debater was trained on the IBM-Rank-30k dataset (Gretz et al., 2019). In order to avoid subjective scales, it was labeled in a relative manner, comparing pairs of arguments (independent of the personal opinion) on 71 controversial topics that were created by crowd-workers. Each argument was annotated by ten different annotators. To derive continuous argument quality scores from the binary annotations, the authors calculate the likelihood of a positive label between 0 and 1, using MACE probability (MACE-P) (Hovy et al., 2013; Habernal and Gurevych, 2016) and Weighted-Average (WA). Both scores inherently incorporate the annotator reliability in the final label. A comparison of the two scoring functions reveals that WA tends to produce a gradual continuous scale, while MACE-P tends to binary labels (i.e., it produces more extreme values close to 0 and 1).

Based on these continuous scores, Gretz et al. (2019) train different models on both WA and MACE-P scores. We focus on the model with the best effectiveness, which is a pre-trained BERT model (Devlin et al., 2019), finetuned in a regression task to predict quality scores given an argument and the corresponding topic. The model is evaluated using Pearson (r) and Spearman (ρ) correlations on the test set (Gretz et al., 2019). Using BERT as contextual language model, Gretz et al. (2019) aim to create an argument quality model that is able to consider quality dimensions such as clarity, relevance and impact of an argument.

Recent works addressing the assessment of argument quality rely on BERT models as well as on "traditional" approaches such as interpreting the sentence lengths (Skitalinskaya et al., 2021; Joshi et al., 2023). An evaluation of the usefulness of large language models for automated argument quality assessment shows a moderate agreement with human annotations, but also demonstrates the potential for improving agreement between annotators (Mirzakhmedova et al., 2024).

3 Reproducing the IBM Model

The original model training process is described in Gretz et al. (2019); we here add missing details and outline how we dealt with these. The authors also referred us to the paper and the original dataset when we asked them for access to the model.

For the pre-trained model on which the IBM model is built, Gretz et al. (2019) link to the official BERT repository,[3] but do not specify which of the various models listed on this page it refers to, except for that is has an output dimensionality of 768. We use the BERT-Base model in the uncased variant, as we assume this is the most frequently used one that matches the description. Following Gretz et al. (2019), we add a linear layer to this pre-trained model and use a sigmoid activation function for the output; the loss is calculated as the mean squared error (MSE). Inspired by Huggingface's BertForSequenceClassification model, we also add a dropout level (with a probability of 0.1), although this is not specified in the paper for the original model, which improves the predictions of our model in preliminary tests. A detailed overview of all training parameters can be found in Table 3 as well as in our public repository (linked in Section 1).

Gretz et al. (2019) do not report the number of training epochs used, nor whether the final model was trained on the WA or MACE-P scores in the dataset (see Section 2). For this reason, we report

[3] https://github.com/google-research/bert

the results for different models trained, based on the WA and MACE-P scores and with a different number of epochs for evaluation on the test set.

The models with the highest scores that achieve similar results in the test set to the original model reported in Gretz et al. (2019) are applied to “external” data as an additional assessment. This data comes from the args.me corpus (Ajjour et al., 2019). This corpus contains argumentative texts on controversial topics that were crawled from various debate portals. In another work of ours (Zelch et al., 2025), we extracted 50 sample texts on different topics from this corpus and split them into sentences (resulting in about 1,100 sentences). While the API of the Project Debater was still available, we predicted the argumentative quality of these sentences using the Debater API to filter out non-argumentative sentences. To evaluate the newly trained models, we compare their predictions on sentences from the args.me corpus with the predictions of the IBM model by calculating the Pearson (r) and Spearman (ρ) correlation coefficients, similar to the evaluation of the original test set.

4 Evaluating the Reproduced Model

We compare the predictions of the reproduced argumentation quality models with the predictins of the IBM model both with the original test dataset and with an “external dataset” that was not used during training.

4.1 Evaluation on the Original Test Set

In a first step, we evaluate our reproduced models similarly to the original IBM model as described in Gretz et al. (2019). Table 1 shows the effectiveness of the reproduced models on the original test data in terms of correlation with the two types of ground truth scores (WA and MACE-P). As the table shows, the original effectiveness on the test set can be achieved within one or two training epochs. With longer training, the model quickly overfits. On average, the models that are trained for two epochs achieve the highest values. For this reason, we use these two models in the following evaluations, one that is trained for two epochs on the MACE-P scores (referred to as MACE-P2), the other that is trained for two epochs on the weighted average scores (referred to as WA2).

4.2 Evaluation on External Data

In addition to the comparison on the original test set, we evaluate the generalization ability of our models on an external dataset. For this, we compare the predictions of the reproduced models and the original IBM model, reporting the deviation between their predictions and fitting a simple linear regression between the original and reproduced models’ scores. As described in Section 3, the external dataset consists of roughly 1,100 sentences from 50 texts on various topics from the args.me corpus (Ajjour et al., 2019) that were labeled while the IBM Project Debater API was still available.

Model	Correlation with Ground-Truth			
	MACE-P		WA	
	r	ρ	r	ρ
Original	0.53	0.52	0.52	0.48
Reproduced				
1 epoch	0.537	0.523	0.532	0.482
2 epochs	0.533	0.522	0.536	0.487
3 epochs	0.485	0.480	0.494	0.441

Table 1: Effectiveness of the reproduced models on the testset, compared to the results reported for the original model by Gretz et al. (2019) in terms of correlation (Pearson’s r and Spearman’s ρ) with the ground-truth.

Model	RMSE	#↘	↘	#↗	↗	r	ρ
MACE-P2	0.359	1132	0.34	1	0.03	0.816	0.823
WA2	0.080	428	0.05	705	0.07	0.901	0.889

Table 2: Correlation between the quality scores of the retrained models and the original IBM scores on args.me arguments; reporting the RMSE, the number of arguments for which the reproduced model predicts a lower score (#↗) or a higher score (#↘), the average distance to the original score for the lower and higher predictions(↗ and ↘), and the Pearson (r) and Spearman (ρ) correlation coefficient.

To evaluate our reproduced models, we compare their predictions on the args.me sentences with the predictions of the IBM model in Table 2, calculating the Pearson r and Spearman ρ correlation coefficient between the predictions. The results show a high correlation between the predictions of the reproduced models and the predictions of the original model as ground truth. The RMSE is low for the model trained on the basis of the weighted average values (WA2), which indicates that the predictions are close to those of the original model. The number of predictions that are lower and higher than the original values is more or less balanced. In contrast, the MACE-P2 model (trained on the MACE-P scores of the dataset) produces consis-

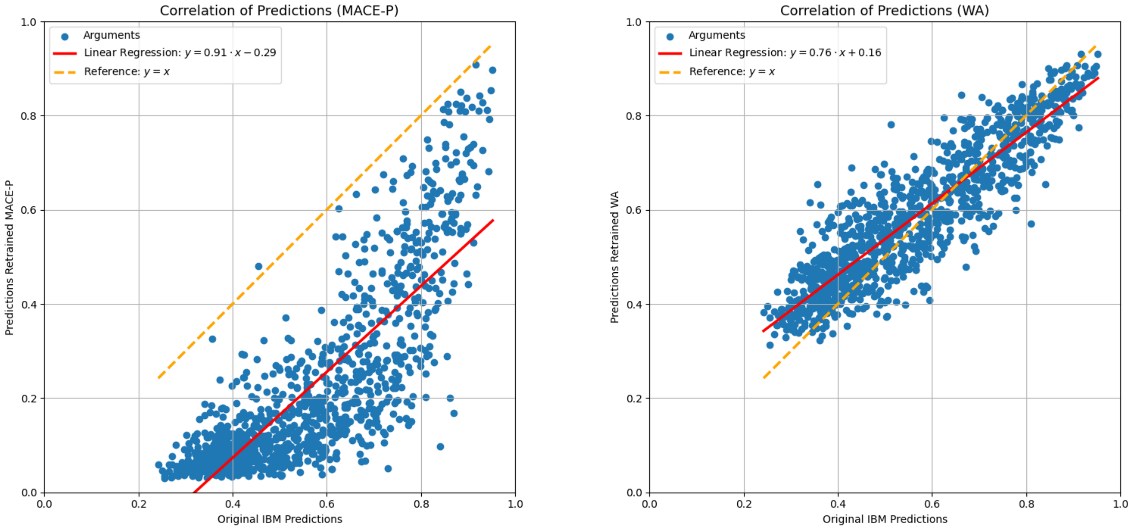

Figure 1: Correlations of the reproduced models' predictions and the original IBM model's predictions on external data (sentences from the args.me corpus). The models are trained on the MACE-P (left) and the WA scores (right).

tently lower scores than the IBM model, which also have a higher deviation (see the higher RMSE). This is consistent with the observation by Gretz et al. (2019) that the models trained on MACE-P scores tend to produce more extreme values close to 0 and 1, while the WA scoring function leads to graduated values.

For the WA2 model, only 12 of 1133 predictions deviate more than 0.2 (but all less than 0.3) from the original IBM predictions. About 20% of the WA2 predictions deviate more than 0.1 from the IBM predictions (223 of 1133), for only 64 of them the difference is greater than 0.15. About half of the WA2 predictions deviate less than 0.05 from the IBM predictions (536 of 1133), 102 of them less than 0.01.

The results in Table 2 are complemented by two scatter plots in Figure 1, which illustrate the correlation between the reproduced models' predictions and the original IBM model for each of the arguments (sentences) from 50 args.me texts. For both models, we show the least squares linear regression ($y = ax + b$) for the given data (red line) and the optimal linear reference (dashed yellow line). For model WA2 (graph on the right), the regression line is close to the optimum, but has a steeper slope (regression coefficients: $a = 0.76$ and $b = 0.16$). The variance is slightly lower at the upper and lower end of the scale. This makes sense, as arguments that are clearly of high or low quality should be easier to identify than arguments of medium quality. Overall, the predictions are roughly in the same range as the original predictions, deviating on average by about 0.05 to 0.07 from the IBM predictions. The scatter plot for the MACE-P2 model (left) looks completely different. The predictions show a strong bias towards lower values and also a significantly higher variance. The regression line has a similar slope to the reference line (regression coefficients: $a = 0.91$ and $b = -0.29$), but is shifted downwards by around 0.3, corresponding to the RMSE in Table 2. The variance of the predictions increases with the improved quality of the arguments, indicating that the models have problems identifying high quality arguments. Based on our evaluation, we therefore assume that the original model was trained on weighted averages with two epochs.

Qualitative Evaluation Table 4 shows exemplary sentences and corresponding quality scores from the second evaluation scenario using args.me texts. The consistently low scores of MACE-P2 are reflected in these examples. There are several cases for which the low predictions are adequate (sentences 6, 7 and 8), however, this cannot be attributed to a good discrimination ability of the model, since most of its predictions are similarly low. The WA2 model predicts similar scores as the IBM model for the examples 2, 3 and 5. In some cases the WA2 predictions appear more reasonable than the IBM scores, such as the higher quality score for sentence 1, as well as the lower score for sentence 6. It is interesting to investigate this in more detail in a follow-up work, to analyze whether one of the two models is consistently better than the other on external data. Both IBM and WA2 seem to have difficulties to recognize non-argumentative sequences, such as the examples 6, 7 and 8 (probably because this kind of data is not present in the training data). However, this is not necessarily problematic as it can be taken into account with an appropriate filtering threshold.

We additionally list the sentences for which the WA2 predictions deviate more than 0.2 from the original IBM predictions in Table 5 (upper half). Interestingly, for these sentences, the reproduced models predictions are all higher than the IBM predictions, except for one (sentence 6). A shared feature of many of these sentences is that they are potentially argumentative for the respective topic when considered together with one or more neighboring sentences—however, they are difficult to interpret without their context. This might also explain the larger deviations in the predictions of the models. The lower half of Table 5 shows the nine sentences with the most similar predictions of WA2 and the IBM model (difference <= 0.001). Several of these sentences with medium scores would not be considered to be very argumentative by humans (e.g., sentence 1132 or 1128), it is interesting that the predicted scores are so similar nevertheless.

5 Conclusion

The paper reports on the reproduction of a model for argument quality prediction that was provided as part of IBM's Project Debater. The original IBM model is not available any longer. With our reproduced models,[4] which follow the training instructions given in Gretz et al. (2019), we achieve similar results on the original test set as reported for the IBM model. On external texts from the args.me corpus, we reach a Pearson's r of 0.9 for the predictions of our best model and the original IBM predictions as ground truth. It is not clear whether this means that the predictions of our reproduced model are worse on the external data, or even better than the predictions of the IBM model. We will address this question in a future work, together with a comparison of our models with more recent approaches.

Limitations

Our reproduced model achieves similar results as the original model on the original test data, and a high correlation with the IBM predictions on foreign data. Nevertheless, the question remains as to where the remaining gap in this correlation and also the partially high variance of the predictions come from. Although we follow the training instructions provided in the paper as good as possible for the reconstruction of the model, some information are not available which might cause small deviations in the training process.

Acknowledgments

This publication has been partially supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU), and by the German Federal Ministry of Education and Research (BMBF) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me Corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of EACL 2023*, pages 957–967.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020. Quantitative Argument Summarization and Beyond: Cross-Domain Key Point Analysis. In *Proceedings of EMNLP 2020*.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI Grand Challenge. In *Proceedings of EMNLP 2021*, pages 267–274.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Lorik Dumani and Ralf Schenkel. 2020. Quality-Aware Ranking of Arguments. In *Proceedings of CIKM 2020*, pages 335–344. ACM.

[4] Repository and models are made available to the public.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis. arXiv/1911.11408.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of ACL 2016*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT 2013*, pages 1120–1130.

Rositsa V. Ivanova, Thomas Huber, and Christina Niklaus. 2024. Let's discuss! Quality Dimensions and Annotated Datasets for Computational Argument Quality Assessment. In *Proceedings of EMNLP 2024*, pages 20749–20779.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. ArgAnalysis35K : A large-scale dataset for Argument Quality Analysis. In *Proceedings of ACL 2023*, pages 13916–13931.

Nailia Mirzakhmedova, Marcel Gohsen, Chia-Hao Chang, and Benno Stein. 2024. Are Large Language Models Reliable Argument Quality Annotators? In *Proceedings of RATIO 2024*, volume 14638 of *Lecture Notes in Computer Science*, pages 129–146.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In *Proceedings of EACL 2021*, pages 1718–1729.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An Autonomous Debating System. *Nat.*, 591(7850):379–384.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5624–5634.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of EACL 2017*, pages 176–187.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an Argument Search Engine for the Web. In *Proceedings of ArgMining 2017*, pages 49–59.

Ines Zelch, Matthias Hagen, Benno Stein, and Johannes Kiesel. 2025. Segmentation of Argumentative Texts by Key Statements for Argument Mining from the Web. In *Proceedings of ArgMining 2025*.

A Appendix

Parameter	Value
Base model	bert-base-uncased
Seed	42
Epochs	2
Batch size	32
Dropout	0.1
Learning rate	2e-5
Epsilon	1e-8
Optimizer	AdamW
Early stopping	no
Added layers	dropout, linear, sigmoid

Table 3: Training parameters for the reproduced argument quality model.

	Sentence	Quality Scores		
		IBM Model	MACE-P2	WA2
(1)	The military is in no obligation to let women into the frontlines just because they hold 95% of the Armies positions, just like why Hooters as no obligation to let men in.	0.52	0.18	0.65
(2)	1) Women already hold just about every kind of post/job in the military and make up a substantial portion of the military My problem with this argument is that it doesn't actually say why the Army should allow women in the frontlines.	0.75	0.48	0.75
(3)	Evidence show the DP is more expensive.	0.45	0.08	0.48
(4)	So there are undoubtedly instances in the past where we have executed an innocent man but did not know so, and still do not know.	0.82	0.29	0.69
(5)	But does this make it right to kill them back?.	0.43	0.06	0.47
(6)	In these cases I made it clear that I could not properly refute my opponent without proper sources.	0.71	0.18	0.56
(7)	Now as for the definition.	0.39	0.14	0.48
(8)	[1] http://www.military.com... [2] http://www.healthline.com... [3] Stuart A. Cohen Israel and Its Army: From Cohesion to Confusion, pg.	0.37	0.24	0.54

Table 4: Example sentences from the args.me corpus on various topics ("We should prohibit women in combat", "We should abolish capital punishment", etc.), along with the quality predictions of the original IBM model and the two reproduced models.

	Sentence	Quality Scores	
		IBM Model	WA2
(1)	Being locked in a single small room in solitary confinement for years on end is certainly not very pleasant.	0.36	0.65
(2)	Just because there are movements for something doesn't mean we should be worried about it.	0.34	0.62
(3)	That's what will soon happen if we can clone, if we can just donor this for that, saving people's lives, people got cured, got strong again, maybe they won't be immortal, but the point is that the increasing of human will soon beyond the balance, causes the disrupt of nature and it's balance, human cloning as you say that is "ethical" can also create tons more of human, also add fuel to that big problem.	0.51	0.78
(4)	(imagine [. . .] they suddenly see that ring on their finger and it sends a flood of guilt through them) In polygamy a very unfair 'status' system will form where only the offspring of the alpha male of the previous generation will be able to compete for the next because all women will think "OOH!	0.43	0.69
(5)	-When people give up all their rights to be protected their is a problem.	0.36	0.61
(6)	It was reported that 0.5% of inmates escaped.	0.81	0.57
(7)	... 2)All the ways that nature preserves that God preserves to help decreasing human population(old age, sickness.	0.32	0.55
(8)	However, when that present is a grenade with the pin pulled out, THEN it becomes immoral.	0.35	0.57
(9)	There are two parts of the act: Giving, and the danger of the grenade" _ So my opponent here believe it is ok to give thee grenade for the present, just don't pull the pin, ok here are the problems with that analogy: 1) How can you give a grenade, a dangerous present to a person whom you loved?	0.33	0.54
(10)	You don't have to learn golf, study it, know the rules and own clubs to be a non-golf player!	0.43	0.64
(11)	More simply, she's protecting rights by protecting rights.	0.48	0.69
(12)	Violating anothers rights does not deprive you of your own: John Stuart Mill is essentially saying the "eye for eye tooth for tooth" concept is right.	0.41	0.61
(1133)	Whether it is or isn't morally correct?	0.47	0.47
(1132)	The goal of debate is to find objective truth.	0.65	0.65
(1131)	First of I would like to say that prostitution is somewhat legal in the U. S. (since only two states allow it, Nevada and Rhode Island).	0.41	0.41
(1130)	Immigration Actually, application for citizenship is still a necessity, as well as a very rigorous INS process which requires applicants to display some sort of evidence of a pre-existing relationship prior to entering the country.	0.64	0.64
(1129)	You might say women have no issue with this, but I will explain.	0.41	0.41
(1128)	My opponent has clearly adopted a strategy based in deception and omission.	0.64	0.64
(1127)	In 2003, Terri Schiavo recovered from a vegetative state that she had been in for 13 years.	0.82	0.82
(1126)	Sure, I'll grant my opponent that there's a correlation; however, we all know that correlation doesn't imply causation, especially considering the maelstrom of recent evidence that I provided in Round 1 suggesting the opposite of Pro's claims.	0.66	0.66
(1125)	Arson is an essential tool in the quest for reform.	0.82	0.82

Table 5: Sentences from the args.me corpus for which the WA2 model's predictions deviate most (> 0.2, upper half) and least (≤ 0.001, lower half) from the original IBM predictions.