LRL_NC at SemEval-2023 Task 4: The Touche23-george-boole Approach for Multi-Label Classification of Human-Values Behind Arguments

Kushagri Tandon, Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, Delhi-110016, India
{kushagri.tandon,niladri.chatterjee}@maths.iitd.ac.in

Abstract

The task ValueEval aims at assigning a subset of possible human value categories underlying a given argument. Values behind arguments are often determinants to evaluate the relevance and importance of decisions in ethical sense, thereby making them essential for argument mining. The work presented here proposes two systems for the same. Both systems use RoBERTa to encode sentences in each document. System1 makes use of features obtained from training models for two auxiliary tasks, whereas System2 combines RoBERTa with topic modeling to get sentence representation. These features are used by a classification head to generate predictions. System1 secured the rank 22 in the official task ranking, achieving the macro F1-score 0.46 on the main dataset. System2 was not a part of official evaluation. Subsequent experiments achieved highest (among the proposed systems) macro F1-scores of 0.48 (System2), 0.31 (ablation on System1) and 0.33 (ablation on System1) on the main dataset, the Nahj al-Balagha dataset, and the New York Times dataset.

1 Introduction

Judgement and opinion of each person are typically governed by a set of principles which is based on a certain value system. Gaining an understanding of human values behind a natural language argument can serve an important role in the field of social sciences, and policy making.

The aim of the current task (Kiesel et al., 2023) is to assign to each argument a set of underlying values, where each argument is given in the form of a Conclusion, Stance and Premise of the argument. With each argument having multiple underlying values, the task effectively boils down to a multi-label classification task, since each argument can have multiple underlying values. English language Touché23-ValueEval Dataset (Mirzakhmedova et al., 2023) has been used for the present challenge.

The labels for the challenge are given as a multilevel taxonomy of 54 human values (Kiesel et al., 2022), called Level 1 labels which are further grouped into 20 human values categories called Level 2 values. The aim of the task is to predict these Level 2 values.

The present work describes two systems: System1 and System2. Both the systems use RoBERTa (Liu et al., 2019) to extract sentence encodings. System1 incorporates Level 1 label information. This system also incorporates linguistic information, learned by predicting the Stance of an argument presented by a Conclusion based on a certain Premise. System2 uses topic modeling to incorporate corpus-level semantic information. System1 was submitted for official evaluation though it was later observed that System2 achieves a higher macro F1-score as compared to System1.

The participating teams were required to make their submissions on the platform TIRA (Fröbe et al., 2023). The official task evaluation platform ranks System1 on the main dataset at the position 22. The proposed systems either outperform or perform at par with the organizer's baseline. But, the systems struggle at capturing low frequency labels.

The paper is organized as follows. Section 2 discusses the task background and Section 3 discusses related past works. Section 4 provides a detailed system overview. The experimental setup and results are given in Section 5 and Section 6, respectively. The paper is concluded with Section 7. The code for the proposed systems have been made available on GitHub.[1]

2 Background

The data made available by the organisers comprises of a main dataset and three supplementary datasets, namely, Zhihu, Nahj al-Balagha and The

[1] https://github.com/KushagriT/SemEval23_ValueEval_TeamLRL_NC

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 136–142
July 13-14, 2023 ©2023 Association for Computational Linguistics

New York Times dataset. The main dataset consists of Train/ Validation/Test counts as 5393/ 1896/ 1576. The task dataset is a collection of 9324 arguments on statements which include religious texts (Nahj al-Balagha), political discussions (Group Discussion Ideas), free-text arguments (IBM-ArgQ-Rank-30kArgs), newspaper articles (The New York Times), community discussions (Zhihu), and democratic discourse (Conference on the Future of Europe). The supplementary Zhihu dataset consists of 100 validation instances. Nahj al-Balagha and The New York Times datasets have 279 and 80 test examples, respectively.

A general instance of the dataset can be interpreted as, Arguing in favor of/against a <Conclusion> by saying <Premise>. An example of such an instance of the dataset, which has been assigned labels or value categories as Security: societal and Universalism: concern, is given in Table 1.

Field	Text
Stance	Against
Conclusion	We should end the use of economic sanctions
Premise	Economic sanctions provide security and ensure that citizens are treated fairly.
Value categories	Security: societal Universalism: concern

Table 1: An instance from the dataset

3 Related Past Works

Recent works have explored role of values and frames in argument mining. A frame is a set of arguments that focus on the same aspect. Ajjour et al. (2019) introduces frame identification towards splitting a set of arguments into a set of non-overlapping frames. They present a fully unsupervised approach to the task which identifies frames using clustering. Kobbe et al. (2020) present models for automatically predicting moral sentiment in debates and explore how moral values in arguments relate to argument quality, stance, and audience reactions. Trautmann (2020) discusses the task of Aspect-Based Argument Mining (ABAM), while Schiller et al. (2021) present a language model for argument generation to generate sentence-level arguments for a given topic, stance, and aspect. Maheshwari et al. (2017) attempt to understand whether individuals in a community possess similar personalities, values and ethical background.

Several recent works also make use of human values. Qiu et al. (2022) present ValueNet, which is a large-scale text dataset for human value modeling. The dataset contains human attitudes on a number of text scenarios. Hendrycks et al. (2023) introduce the ETHICS dataset, a new benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality. They also analyse the ability of language models to predict basic human ethical judgements. Ammanabrolu et al. (2022) create interactive chat agents which act in alignment with socially beneficial norms and values. Liu et al. (2023) introduce a new learning paradigm Second Thoughts that can make current language models aware of the human value alignment.

4 System Overview

This section discusses System1 and System2 as mentioned in Section 1 . The first step in the systems is to generate an encoding for each instance using RoBERTa (Liu et al., 2019).

Figure 1 and Figure 2 give the architecture of the System1 and System2, respectively.

4.1 System1

The core idea for this system is to combine the information of factors that classify each argument into Level 1 labels with the information which can predict the stance indicated by the conclusion drawn from a particular premise. To generate sentence encodings, the RoBERTa model has been fine-tuned for Masked Language Modeling (MLM) task on the main task dataset. Each instance consists of three entries, namely, Conclusion, Stance and Premise, and input text is represented as, <Conclusion> + ‘ </s> ’ + <Stance> + ‘ </s> ’ + <Premise>, where </s> is the separator token. The RoBERTa layer of this fine-tuned model is called MyRoBERTa in the rest of the paper.

This system consists of two sub-models trained for related auxiliary tasks.

- The first sub-model, called Model_Level1, is essentially the MyRoBERTa model fine-tuned for multi-label classification on the set of 54 Level 1 labels from the human values taxonomy. It takes as input <Conclusion> + ‘ </s> ’ + <Premise>. This model is trained using Binary Cross Entropy loss which is preceeded by

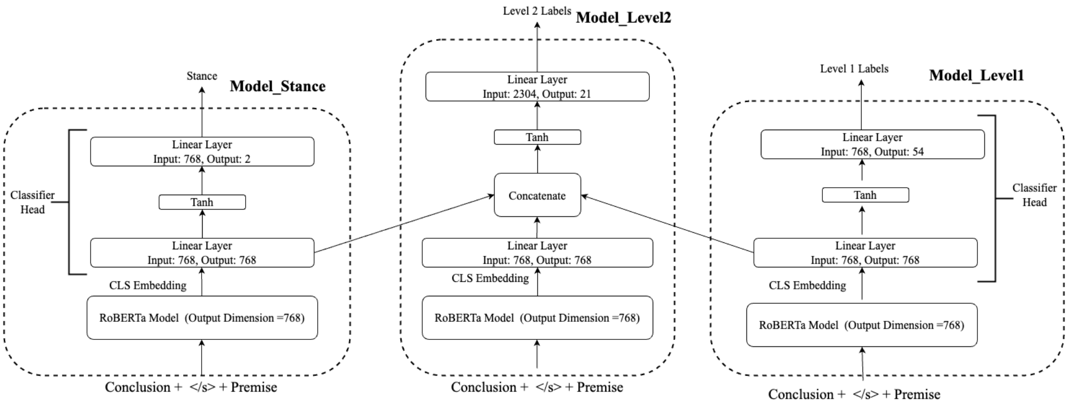

Figure 1: **System1**: It consists of two sub-models, Model_Level1 and Model_Stance. Model_Level1 consists of the MyRoBERTa layer, followed by a classification head to generate the label subset. The classification head consists of a dense neural network layer with input and output dimensions as embedding size of the RoBERTa model. It takes as input the CLS token embedding from the MyRoBERTa layer. The output from this layer is applied with Tanh activation, and is sent as input to a dense neural network layer with output dimension as the target space size corresponding to the Level 1 labels i.e., 54. Model_Stance consists of a MyRoBERTa layer to generate sentence encoding, followed by a classification head, to generate binary predictions. The model for multi-label classification on the set of Level 2 labels of size 20, is called Model_Level2.

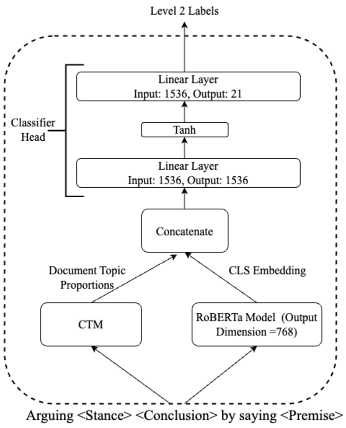
Level 2 Labels
Classifier Head
Linear Layer
Input: 1536, Output: 21
Tanh
Linear Layer
Input: 1536, Output: 1536
Concatenate
Document Topic Proportions
CLS Embedding
CTM
RoBERTa Model (Output Dimension =768)
Arguing <Stance> <Conclusion> by saying <Premise>

Figure 2: **System2**: The first layer of this system is RoBERTa and the corresponding CLS embeddings from this layer are concatenated with the K-dimensional text topic proportions from Correlated Topic Model.This is input to a classification head, of architecture as described in case of Model_Level1 of System1.

application of sigmoid layer on the output.[2].

- The second sub-model, called Model_Stance, is essentially a MyRoBERTa model fine-tuned for binary classification on stance of the argument given the input text, <Conclusion> + ‘ </s> ’ + <Premise>. This model is trained using Cross Entropy loss.

The model for multi-label classification on the set of Level 2 labels of size 20, is called Model_Level2. The first layer of which is MyRoBERTa[3]. The CLS embedding from this layer is input to a dense neural network layer with input and output dimensions as embedding size of the RoBERTa model. The output from this layer is concatenated with the output from the first dense layer of the classification head of each of the two auxiliary models trained above. Tanh activation is applied to this concatenated embedding and input to another dense neural network layer with output dimension as the target space size corresponding to the number of labels. This model is trained using Binary Cross Entropy loss which is preceded by application of sigmoid layer on the output.

While training the three models Model_Level1, Model_Stance, and Model_Level2, all the parameters of the MyRoBERTa layer are frozen, except the parameters corresponding to the last and second-to-last encoder layer.

4.2 System2

This approach combines topic modeling with RoBERTa. Topic modeling takes into account the general semantic properties of the entire corpus. Analysis of the corpus reveals that often the arguments are centered around certain human values and tend to involve some specific words. Moreover, they exhibit a significant correlation between the semantically coherent units or topics that are represented by the data. Hence we train a Correlated

[2] BCEWithLogitsLoss in huggingface transformers

[3] The pooling layer is not added to this RoBERTa model.

Topic Model (CTM) (Lafferty and Blei, 2005) on the corpus, where each instance is transformed as 'Arguing '+ <Stance> + <Conclusion> + ' by saying ' + <Premise>. CTM produces a K-dimensional vector which represents the topical composition of that argument.

The first layer of this system is RoBERTa[4] from pretrained roberta-base model. The instance 'Arguing '+ <Stance> + <Conclusion> + ' by saying ' + <Premise> is input to this model, and the corresponding CLS embeddings are concatenated with the K-dimensional text topic proportions. This serves as an encoding for each instance. This encoding is passed as input to a classification head, of architecture as described in case of Model_Level1 of System1. This model is trained for multi-label classification on the set of Level 2 labels of size 20, using Binary Cross Entropy loss on the output, where a sigmoid layer has been applied to the output to convert the values to probabilities.

5 Experimental Setup

The experiments were carried on Google Colaboratory in Python 3.8.10 with Nvidia Tesla P100 GPU. PyTorch (Paszke et al., 2019), and Huggingface Transformers (Wolf et al., 2020) were used as the key frameworks for the experiments.

Training is carried on the `Train + Validation` subsets of the main dataset. For validation, the `Validation` subset of the supplementary Zhihu dataset, has been used in order to make the system more robust.

5.1 Description of MyRoBERTa

For finetuning RoBERTa for MLM task the following settings were followed.

- The roberta-base is fine-tuned on the combined `Train + Validation` subsets of the main dataset.

- The text was tokenized using RobertaTokenizerFast from Huggingface Transformers and RoBERTa special tokens were added. The data was prepared by masking tokens in the input with probability 0.15. The instances were padded to the maximum sequence length in the batch. Truncation was done at a sequence length of 256. The model was implemented using RobertaForMaskedLM from Huggingface Transformers.

- The model was fine-tuned with batch size 4, for 6 training epochs and learning rate 5e-5.

5.2 Description of System1

To train Model_Level1 and Model_Stance the following steps were used.

- The data was prepared and batched by sorting the input text, <Conclusion> + ' </s> ' + <Premise> according to text length, padding to maximum document length in the batch, and truncating at a maximum sequence length of 256. The text was tokenized using RobertaTokenizerFast, and RoBERTa special tokens were added.

- The optimizer used is Adam with weight decay (Kingma and Ba, 2014; Loshchilov and Hutter, 2017).

- The task of Model_Level1 is multi-label classification with 54 labels, and the task of Model_Stance is binary classification with 2 labels.

- The hyperparameters for the model are given in Appendix.

The two sub-models are not further trained as a part of the architecture of Model_Level2. The model settings for both the sub-models and Model_Level2 are given in the Appendix. The final model state for Model_Level2 was selected according to the highest macro F1-score on the development dataset.

5.3 Description of System2

The Correlated Topic Model (CTM) was trained using using tomotopy package[5] in Python. The text was preprocessed using the simple_preprocess in Gensim (Řehůřek and Sojka, 2010). The standard English stopwords were removed using the NLTK toolkit[6]. The words having POS (part-of-speech) tags as NOUN, ADJ (adjective), VERB and ADV (adverb) were lemmatized. But proper nouns were retained without change. The POS tagging and lemmatization were done using Spacy[7] framework in Python.

The settings for training the CTM model are given in Appendix. For training the main model,

[4] The pooling layer is not added to this RoBERTa model.

[5] https://bab2min.github.io/tomotopy/v/en/

[6] https://www.nltk.org/

[7] https://spacy.io/

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best per category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
System1	.46	.49	.61	.05	.20	.61	.28	.47	.23	.74	**.61**	.49	.49	.27	**.19**	**.53**	.14	.71	.77	.34	.41
Ablation 1.1 *	.47	.51	.61	**.18**	.18	.60	**.31**	**.51**	.28	**.75**	.60	.52	.49	.46	.17	.50	.16	.72	.79	**.40**	.37
Ablation 1.2 *	**.48**	.52	.62	**.18**	.18	**.62**	.25	**.51**	**.29**	.74	**.61**	.52	**.54**	.49	.23	.52	.18	.72	**.82**	.34	**.46**
Ablation 1.3 *	.46	.51	.62	.05	.12	.60	.29	.45	.17	**.75**	**.61**	**.54**	.49	**.54**	.12	**.53**	.17	.73	.80	.36	.37
System2 *	**.48**	**.53**	**.63**	.11	**.29**	.61	**.31**	.48	.14	.72	**.61**	**.54**	.49	.48	.14	**.53**	**.24**	**.75**	.78	**.40**	**.46**
Nahj al-Balagha																					
Best per category	.48	.18	.49	.50	.67	.66	.29	.33	.62	.51	.37	.55	.36	.27	.33	.41	.38	.33	.67	.20	.44
Best approach	.40	.13	.49	.40	.50	.65	.25	.00	.58	.50	.30	.51	.28	.24	.29	.33	.38	.26	.67	.00	.36
BERT	.28	.14	.09	.00	.67	.41	.00	.00	.28	.28	.23	.38	.18	.15	.17	.35	.22	.21	.00	.20	.35
1-Baseline	.13	.04	.09	.01	.03	.41	.04	.03	.23	.38	.06	.18	.13	.06	.13	.17	.12	.12	.01	.04	.14
System1	.27	.07	**.30**	.29	.22	.55	.18	.00	.18	.45	.21	.29	**.26**	**.27**	**.27**	.29	.30	.21	.00	.10	**.32**
Ablation 1.1 *	.27	**.15**	.28	.25	.36	.61	.00	**.25**	.31	.44	**.35**	.34	.24	.08	.14	.33	**.33**	**.28**	.00	.00	.30
Ablation 1.2 *	**.31**	.13	.28	**.40**	.44	.60	.00	**.25**	.33	.39	.33	.38	**.26**	.17	.14	**.36**	.30	.25	.00	**.13**	.31
Ablation 1.3 *	.30	.10	.29	**.40**	**.50**	.58	.00	.00	.20	**.47**	.28	.47	.25	.12	.26	.34	.30	.34	.00	.10	.34
System2 *	.28	.14	.22	.00	.40	**.65**	**.22**	.00	**.46**	.43	.26	**.49**	.23	.09	.10	.31	.30	.27	.00	.08	.30
New York Times																					
Best per category	.47	.50	.22	-	.03	.54	.40	-	.50	.59	.52	-	.33	1.0	.57	.33	.40	.62	1.0	.03	.46
Best approach	.34	.22	.22	-	.00	.48	.40	-	.00	.53	.44	-	.18	1.0	.20	.12	.29	.55	.33	.00	.36
BERT	.24	.00	.00	-	.00	.29	.00	-	.00	.53	.43	-	.00	.00	.57	.26	.27	.36	.50	.00	.32
1-Baseline	.15	.05	.03	-	.03	.28	.03	-	.05	.51	.20	-	.07	.03	.12	.12	.26	.24	.03	.03	.33
System1 *	.26	.29	.00	-	.00	.17	.00	-	.00	.54	.39	-	**.46**	.00	**.29**	.25	.35	.50	.50	.00	.38
Ablation 1.1 *	**.33**	.33	.00	-	.00	.25	.00	-	.25	.57	**.46**	-	.20	**1.0**	.00	.30	**.43**	**.56**	**.67**	.00	.40
Ablation 1.2 *	.24	.25	.00	-	.00	.32	.00	-	.29	.52	.42	-	.15	.00	.00	**.38**	.38	.42	.40	.00	.35
Ablation 1.3 *	.32	.29	.00	-	.00	**.37**	.00	-	**.33**	**.61**	.30	-	.13	**1.0**	.00	.28	.41	.48	.50	.00	**.45**
System2 *	.27	**.40**	**.18**	-	.00	.36	.00	-	.00	.55	.52	-	.25	.00	.00	.15	.23	.43	**.67**	.00	.42

Table 2: Achieved F_1-score of team george-boole per test dataset. Approaches marked with * were not part of the official evaluation. Approaches in gray are an ensemble using the best participant approach for each individual category; the best participant approach; and the organizer's BERT and 1-Baseline. The best scores for each category, among the proposed systems are boldfaced

the data was prepared by considering the input text, 'Arguing '+ <Stance> + <Conclusion> + ' by saying ' + <Premise>, and sorted according to text length. The instances in a batch were padded to maximum document length in the batch, and truncation was done at a maximum sequence length of 256. The final model state was selected according to the highest macro F1-score on the development dataset. The model settings used for training the main multi-label classification model are given in Appendix.

6 Results

Ablation studies on System1 were conducted as follows. The results are given in Table 2.

- Ablation 1.1: System1 without Model_Level1
- Ablation 1.2: System1 without Model_Stance
- Ablation 1.3: System1 without Model_Level1 and Model_Stance

In each of three test sets the submitted systems either outperform or perform at par with the organizer's BERT and 1-Baseline. Among the proposed systems, System2 performs the best on the Main test set. Whereas, Ablation 1.2 performs best in case of Nahj al-Balagha test set. In case of New York Times test set, the Ablation 1.1 performs almost at par with the best participant model on the official task leaderboard based on the the macro-F1 score. This can be attributed to the use of the supplementary Zhihu subset for validation, improving the robustness of the system. It is observed that use of topic modeling improves the performance of the system, in general.

In our experiments, on the main dataset, the F1-score is highest for eleven labels, for System2. Similarly, on the Nahj al-Balagha dataset, for five labels, the the F1-score is highest for Ablation 1.2. For this dataset, System2 performs at par with the BERT baseline. On the New York Times dataset, for five labels, the F1-score is highest for Ablation 1.1.

7 Conclusion

The current work proposes two systems for the task of value identification in arguments. This is broadly a multi-label classification task. The proposed systems use RoBERTa to get sentence encodings for each document. System2, combines RoBERTa with topic modeling to get sentence representation. This is followed by a classification head to generate the label subsets. The other system, namely, System1, also makes use of features obtained from training models for auxiliary tasks. The auxiliary tasks, are that of prediction of Level 1 values from the human-value taxonomy given an argument, and stance prediction given conclusion and premise of an argument.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

John Lafferty and David Blei. 2005. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony X Liu, and Soroush Vosoughi. 2023. Second thoughts are best: Learning to re-align with human values from text edits.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Tushar Maheshwari, Aishwarya Reganti, Upendra Kumar, Tanmoy Chakraborty, and Amitava Das. 2017. Semantic interpretation of social network communities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

All transformer-based models were trained using Trainer from Huggingface Transformers.

For finetuning RoBERTa for MLM task, the dataset was batched using, DataCollatorForLanguageModeling.

The model settings for both the sub-models and Model_Level2 are given in Table 3. These models are implemented using RobertaForSequenceClassification from Huggingface Transformers, with pretrained model as MyRoBERTa.

Setting	Model_Level1 Model_Stance	Model_Level2
Learning Rate	5e-4	5e-5
Weight Decay	0.01	0.01
Optimizer	AdamW	AdamW
Epochs	25	20
Batch Size	8	8

Table 3: Model Settings for System1

The settings for training CTM model are given in Table 4. The settings for train the main multilabel classification model in System2, are given in Table 5.

Setting	Value
Term Weighing Scheme	IDF
minimum collection frequency	5
k	40
Burn-in Samples	5
Iterations	200

Table 4: CTM Model Settings

Setting	Value
Learning Rate	3e-5
Weight Decay	0.001
Optimizer	AdamW
Batch Size	4
Epochs	6

Table 5: Model Settings for System2