Group9 et al., at SemEval-2023 Task 4: Value Category Prediction using Contrastive Learning and Textual Entailment

Arnav Agarwal, Aayush Ranjan, Udit IIITD, Jyotirmaya Singh
IIIT Delhi

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

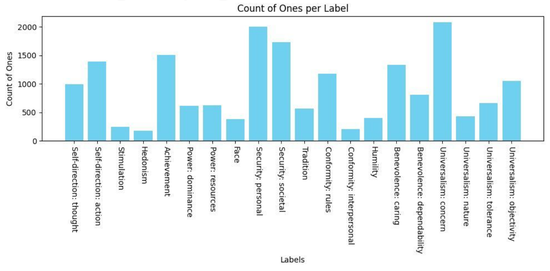
Introduction

Problem Statement: Given a textual argument containing premise,stance and conclusion predict whether or not an argument draws on a given value category.

Challenge: Arguments given by a speaker are governed by their value system. Consider for instance the arguments:

1. Capital punishment has been proven ineffective in reducing crime rates. (Universalism:Objectivity)
2. There are too many historical cases where innocent people have been put to death. (Universalism: Concern)

Yet, they could be useful in personalisation of chatbots, in pedagogy and targeted political and advertising campaign.



Dataset: We used the Touche23-ValueEval Dataset containing 9324 arguments from 6 diverse sources.

Task Modelling: Multi-Label classification where each triplet of Premise,Stance, Conclusion could draw from one or more of 20 value categories.
x_i=<P,S,C>: Premise, Stance, Conclusion triplet.
(l_1,...l_{20})->20 value categories.
f: x_i->{l_j | j∈J} J={1,2,3,...20}

Contributions

- Proposed a novel architecture that combines textual-entailment and contrastive learning strategies. Obtained an F1 that beats F1 by winner of SemEval 2023.
- Developed another model to try and explain predictions generated by us.
- Proposed a novel architecture that combines the classification task with a generation task as a co-learning approach.

Methodology

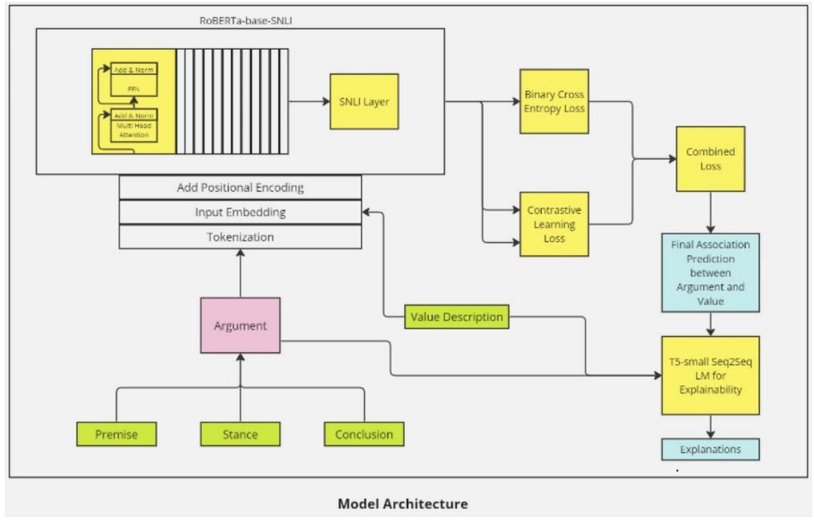
Following on the idea proposed by Kierkegaad[1] et al., we decide to model the problem as a textual entailment task. This has the following motivation:

1. Increase number of samples for training.
2. Make the model label-aware.
3. Allows us to use pre-trained NLI models and improve results by modelling the task exactly like the training task for these models.

We treat the combined argument, stance, and conclusion as the premise and the value description as the hypothesis. We model this as binary classification and fine-tune the pepa/roberta-base-snli model for our specific task.

Further we decide to use ContrastiveLoss in addition to our standard classification loss as proposed by TM Scalcon. This allows the model to learn robust and discriminative embeddings by maximizing the distance between dissimilar examples and minising distance of similar ones.

By learning relationships beyond those exhibited by input-label pairs the model will be able to learn a much richer embedding and make it robust to unseen data as well as labels with few samples.



Model Architecture

Methodology

Experimental Setup:
Model- pepa/roberta-base-snli
Epochs- 3
Optimiser- AdamW
Learning Rate- 2e-5
Loss- Contrastive Loss + BCE

In addition to the above architecture we also tried multiple configurations: 1) Entailment only, 2) Contrastive Only 3) Ensemble 4) Generation of Conclusions

Explainability: We further use a separate model to generate explanations for our predicted labels.

Argument	Value Category	Generated Explanation
Subsidizing vocational education will lead to it becoming as esoteric and non-productive as liberal arts education.	Humility	Avoid upsetting or annoying others, being tactful to others, showing courtesy, being polite, resisting temptation, respecting elders.
It should be legal so that police can get all the drug dealers off the streets	Achievement	Being ambitious, successful and being admired for achievements and skills. Demonstrating competence according to social standards or in competition. Important dimension of achievement is that it is perceived within the social standards, and according to the rules of engagement, unlike

We used a t5 model fine tuned on the training dataset to produce the explanations. While the fluency of generated explanations was found to be high, the model showed divergence in adequacy of context. Cohen's Kappa for inter-annotator agreement on adequacy for the two human annotators was found to be 0.7736.

Results and Findings

Model	Val Set				Test Set			
	Acc	F1	Prec	Recall	Acc	F1	Prec	Recall
Entailment	0.74	0.45	0.31	0.55	0.71	0.44	0.29	0.59
Contrastive	**0.78**	0.47	0.35	0.59	0.75	0.46	0.33	0.56
Ensemble	0.63	0.41	-	-	0.61	0.40	-	-
Conclusion Generation	0.77	0.26	-	-	**0.69**	0.25	-	-
Entailment+ Contrastive	0.73	**0.58**	**0.63**	**0.64**	0.73	**0.56**	**0.63**	**0.64**

- The Entailment+Contrastive model when tested on the main test set produces a macro F1 of 0.5681 which is slightly better than the previous best F1 score of 0.561.
- On Nahj-al-Balagha we obtain an F1 score of 0.46 which is a huge jump from previous best of 0.36, showing the improved robustness of our model.
- While F1 on rare labels is slightly poor, we do notice an improvement as compared to other models and previous submissions. This shows our model generalised well to labels with few instances.

Conclusion and Future Directions

1. Achieved high F1 scores by enhancing RoBERTa-base-SNLI model with contrastive loss for binary classification of value categories within arguments.
1. Developed secondary model for explanation generation, with ongoing refinement efforts focused on improving explanation adequacy and computational efficiency.
2. Future plans include exploring data augmentation, reinforcement learning, and local explanation generation to enhance model adaptability, accuracy, transparency, and utility in real-world applications.

References

1) I. T. Cepeda, A. Pauli, and I. Assent, "Sren kierkegaard at semeval-2023 task 4: Label-aware text classification using natural language inference
2) M. M. Oskuee, M. Rahgouy, H. B. Giglou, and C. D. Seals, "Tm scanlon at semeval-2023 task 4: Leveraging pretrained language models for human value argument mining with contrastive learning