Quintilian at SemEval-2023 Task 4: ValueEval: Identification of Human Values behind Arguments

Ajay Narasimha Mopidevi
Ajay.Mopidevi@colorado.edu
University of Colorado, Boulder

Hemanth Chenna
Hemanth.Chenna@colorado.edu
University of Colorado, Boulder

Abstract

In this paper, we initially propose our approaches to model the ValueEval task, using Natural Language Inference methods. In the later sections, we explain in detail about our proposed models for multi-label classification. We have proposed a hybrid architecture which leverages the commonality between the classes to better perform at a multi-label classification tasks. Our proposed architecture outperforms all of the baseline models.

1 Introduction

Everyone has a perspective of how they approach a problem and make a decision, from the simplest decisions such as choosing to lend a pen to a friend to decisions that affect their life and everyone around them. These decisions are made consciously based on some of the values that they strongly believe in. The task 4 of SemEval-2023 - ValueEval, is to understand which of these human values form the basis for someone's decision in a textual argument.

Kiesel et al. (2022a) presents the task of understanding human values as a Natural Language Processing problem, by proposing an annotated dataset of textual arguments, with the labels being the human values that would be drawn to make that decision. Each argument in this dataset is provided with a Premise, Conclusion and Stance. Lets consider this example argument from the dataset with premise "marriage is the ultimate commitment to someone, if people want it they should be allowed", the conclusion "We should abandon marriage" and the stance "against". The task is to figure out why someone takes a particular stance (in favor of/ against) for the conclusion, given the premise - "What human values led to someone taking this particular stance?"

The task studies the (often implicit) human values behind natural language arguments, such as to have freedom of thought or to be broadminded. Values are commonly accepted answers to why some option is desirable in the ethical sense and are thus essential both in real-world argumentation and theoretical argumentation frameworks. However, their large variety has been a major obstacle to modeling them in argument mining.

In their dataset, Kiesel et al. (2022a) also provided each argument with a multi-level taxonomy of human values in the form of labels that are closely aligned with psychological research. The 54 labels in level 1 are grouped into 20 labels in level 2, which are further grouped into 4 labels in level 3, and finally grouped into two separate levels 4a and 4b each with 2 labels. For each argument, level 1 labels provides finer details about which human values are inferred for making the decision but as the level increases, the labels generalize into a smaller set of linked human values which are computationally much easier, but provide much less granularity in the final classification of values.

Each of argument in the dataset may draw from multiple labels at each level. In this paper, we try to classify the labels for level 2, which provides a certain degree of finer details about the arguments but also generalizes some of the similar or linked human values to make it computationally easier.

2 Related Works

This classification task becomes much more challenging as it is a multi-label classification problem. In classification problems, deep learning architectures try to update their layer weights to emphasize the output of softmax layer of the correct class by making it closer to 1, while also making the outputs of other classes closer to 0. In a multi-label classification problem, that's not possible as each datapoint may have multiple labels. Tsoumakas and Katakis (2007) provides a list of approaches used to tackle multi-label classification problems and provides a comparison of the performance of these approaches. They showed that Boutell et al. (2004) PT3 transformation provides better results

compared to other transformations. Considering that level 2 has 20 labels, a PT3 transformation generates many more classes, and also reduces the number of samples per class, making the data very sparse. In such tasks with more labels, Lauser and Hotho (2003) PT4 transformation is preferred as it uses L binary classifiers, with each binary classifier predicting 1/0 for each class.

Zhang and Zhou (2013) mentions that extracting high-level relations among the classes can improve the performance of multi-label classification tasks. Ji et al. (2008) tries to illustrate relations on how a class influences the other classes, while Read et al. (2008) establishes relations among a random subset of classes.

Along with the dataset, Kiesel et al. (2022a) provides a few baseline models for comparison. We have the following 3 baseline models: a 20-label classifier network that uses contextual embeddings from a pretrained BERT model, an SVM model and a 1-Baseline which predicts the label 1 for all the classes. While their initial results look promising, they only considered the premise for their classification, ignoring the conclusion and stance. As they ignore this information, these models don't properly infer which human values are the reasons behind making a stance in an argument.

3 Dataset Analysis

As described in the Introduction, the provided training, validation and test datasets contain textual arguments of two sentences - a premise and a conclusion - with a stance, and based on the level, one row of 1s and 0s for each datapoint that represent if the reason for choosing that stance draws from that particular human value. We are focusing on classifying Level 2 labels, so there will be 20 such binary values in each row.

For level 2, the training dataset contains 4240 arguments with labels, the validation dataset contains 277 arguments with labels and the test dataset contains 753 datapoints of arguments and labels. For the SemEval-2023 workshop, the organizers have another test dataset of just arguments which will be used for grading submitted test runs. Since the dataset contains such a small number of datapoints, expanding this dataset to try and classify using PT3 transformation would lead the small dataset to be even more sparse with possibly 0 datapoints for many PT3 transformed classes. This is further justification for us to prefer PT4 transformation over PT3 transformation.

Label Cardinality(LC) is computed as the average of the true labels in the input data. Label Density(LD) is similar to Label Cardinality, but is also divided by the total number of labels in the input data. (Venkatesan and Er, 2014). We use this information to draw further insights about our dataset. These values help us identify how sparse the labels are for each datapoint in our dataset. For the training dataset, we have observed the LC and LD values in Table 1

From these LC and LD values, we observe that we can make groupings in our hybrid model into 5 groups of 4 as intermediate classes, before classifying them into their respective classes. This grouping has been explained in further detail in the Hybrid model subsection of the Methodology section.

Label Cardinality	3.4607
Label Density	0.1730

Table 1: Label Cardinality and Label Density for training and validation datasets

4 Methodology

We model this problem as a Natural Language Inference (NLI) model, to predict the human values that led to the stance being taken based on the premise and the conclusion. NLI typically has only two inputs i.e premise and conclusion and the output is stance. We modified our conclusion to have the stance followed by the actual conclusion. From here, the use of the term conclusion includes the stance appended to the beginning of the original conclusion. The premise and this new conclusion pair are jointly encoded using BERT to obtain the contextual information between the premise and the conclusion along with the contextual information of the premise tokens and the contextual information of the conclusion tokens. This contextual information in the form of embeddings are transferred to the classification network.

In our paper, we propose three different classification models:

- L-Label Classifier Architecture
- L-Binary Classifiers Architecture
- Hybrid Architecture

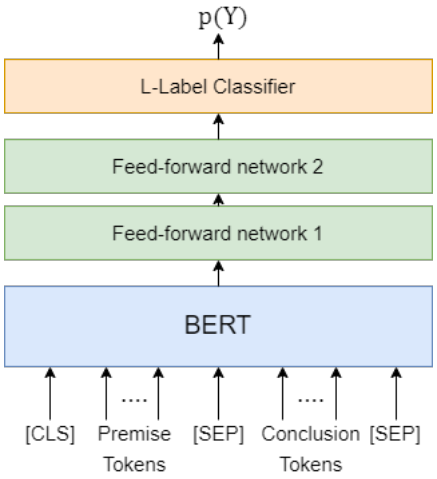

Figure 1: L-Label Classifier Architecture

4.1 L-Label Classifier Architecture

The embeddings from BERT pretrained model Devlin et al. (2018) are forwarded to the feed forward layers and finally a classifier layer. The L-label classifier as shown in Figure 1, predicts the probabilities for each of the L classes (L=20) and assigns the labels with a probability above a certain threshold to be 1. For our model, we used a threshold of 0.5 and we use Cross Entropy loss as our loss function.

4.2 L-Binary Classifiers Architecture

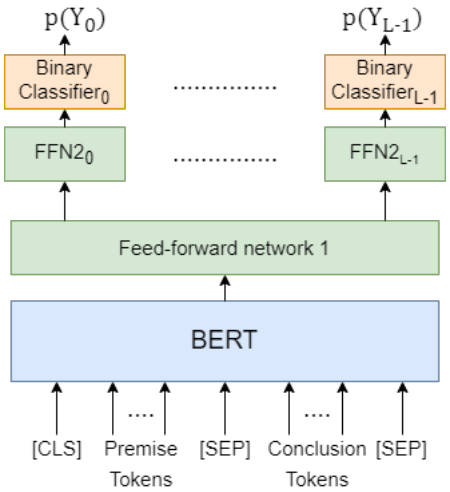

Figure 2: L-Binary Classifiers Architecture

Lauser and Hotho (2003) proposes a method which transforms the multi-label classification task as set of L- binary classification tasks. Using this approach, we modified our network to an L-binary classifiers architecture, where each binary classifier predicts 0/1 of a particular label as displayed in Figure 2. For each of these binary classifiers, the loss function we use is Binary Cross-Entropy loss.

4.3 Hybrid Architecture

The binary classifiers in our previous model are trained independently for each label. This loses valuable information stored in the commonalities between the different classes. As described in our analysis of the dataset, the label cardinality of the dataset is ∼3.4 i.e. each stance for an argument draws from approximately 3 or 4 human values. This confirms the insight we can draw from the analysis by Zhang and Zhou (2013) - exploring the high-level relations between the classes, we can make better predictions of the multiple labels.

Instead of approaching a random grouping, we group a subset of labels in each category based on the similarities. For finding these similarities, we extract the embeddings of the classes and perform a k-Means clustering. We used a constrained k-Means algorithm proposed by Bradley et al. (2000) to exactly model each group with 4 labels. This number is chosen as it close to the label cardinality.

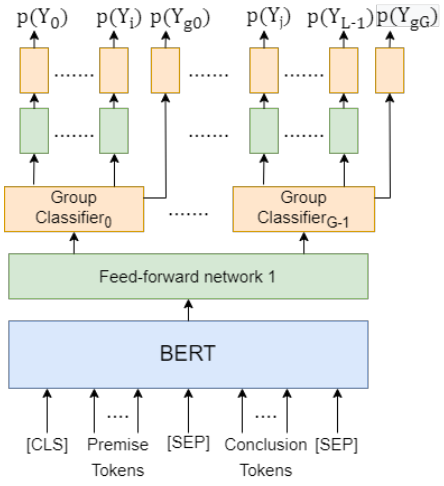

Figure 3: Hybrid Architecture

The output of each group classifier is forwarded to the hidden layers and subsequently the L binary classifiers for each class, and also to a separate classifier layer to predict whether the final set of labels belong to that group. We need to create extra labels for these groups for loss calculation, by evaluating its member labels in that group. We have experimented with the standard Cross Entropy loss and Hamming loss and combined these losses to train our model.

5 Results

We compare each of our approaches using the F_1 score, Precision (P), Recall (R) and Accuracy (A) against the baselines provided in Kiesel et al. (2022b). These metrics are individually calculated for each of the labels and finally scores are presented as an average across all the labels. As it is a multi-label classification and the dataset is not a balanced dataset, it is better to consider F_1 scores. We trained all of our models for 200 epochs each with a variable learning rate, linearly rising to peak($1e^{-4}$) at 40 epochs and then linearly declining back to 0 by the 200th epoch.

Model	P	R	F_1	A
1-Baseline	0.18	1.0	0.28	0.18
SVM	0.30	0.30	0.3	0.77
BERT	**0.39**	0.30	0.34	**0.84**
L-label classifier	0.29	**0.48**	0.36	0.76
L-Binary classifiers	0.3	0.45	0.35	0.77
Hybrid+ CE loss	0.32	0.42	0.39	0.77
Hybrid+ CE + HD loss	0.33	0.43	**0.40**	0.77

Table 2: Performance of our approaches compared to baseline

For our hybrid architecture, we have grouped the classes as follows:

- 'Achievement', 'Face', 'Power: dominance', 'Power: resources'
- 'Benevolence: caring', 'Benevolence: dependability', 'Humility', 'Universalism: concern'
- 'Stimulation', 'Tradition', 'Self-direction: action', 'Self-direction: thought'
- 'Conformity: interpersonal', 'Conformity: rules', 'Security: personal', 'Security: societal'
- 'Hedonism, Universalism: nature', 'Universalism: objectivity', 'Universalism: tolerance'

All our proposed architectures perform better than the baseline models according to the F_1 score. Our hybrid architecture shows a significant improvement in F_1 scores compared to our L-label classifier model, the L-binary classifiers model and all the baselines. This is mainly achieved due to the grouping of similar labels, which adds more detail about the possible relations between labels for a given argument.

The L-binary classifiers F_1 score is slightly lesser than the L-label classifier as each of the labels are computed without any knowledge of the other labels prediction. This model completely loses out the common information. This further strengthens our hypothesis to group the classes and extract the relationship between the labels.

We experimented with multiple loss functions for our hybrid architecture. *Hybrid + CE loss* only uses cross entropy loss functions, while *Hybrid + CE + HD loss* uses both cross entropy loss and Hamming Distance(HD) loss. Although adding HD loss boosted the F1 score, the improvement is very negligible.

From the Figure 4, we can see that both our Hybrid models perform as good or better than all of the other models for each of the individual labels as well. In some classes, the hybrid models vastly outperform both the other models.

6 Link to Github Repo

https://github.com/HemanthCU/NLP_SharedTask_Task_4

7 Conclusion

In this paper, we presented our approach for the ValueEval task of SemEval 2023. By modeling our problem as an NLI problem, we are able efficiently utilize all the components of the input argument i.e premise, conclusion and stance to predict the reason for the stance. This approach has made all our models to perform better than the provided baseline models.

Also extracting the similarities between the classes and grouping them has significantly increased the F_1 score. As the similar labels are grouped together, this reduced the complexity in multi-label classification, leading to better performance.

8 Future Work

We expected that our grouping would reduce the label cardinality to close to 1.0, but the observed value for the groups is $\sim$2.02. Although this is less

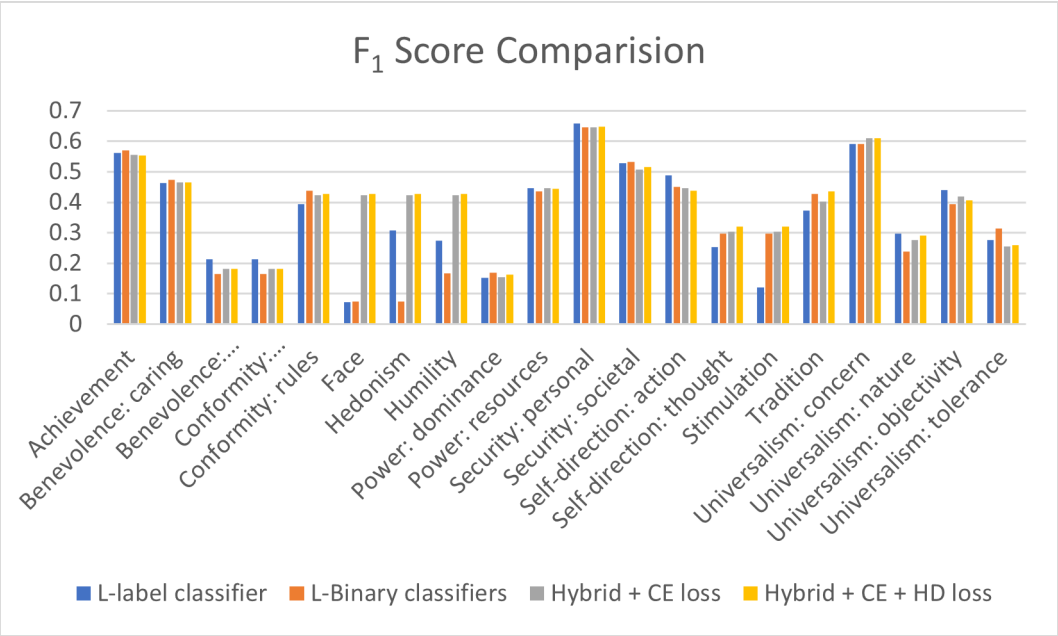

Figure 4: Comparison of individual test set F_1 scores for each of the labels by the different models we have trained and tested

compared to original value (3.4), the groups are also multi-label. We would like to extend our work to model the the group prediction as a multi-class task, while the labels inside the group need to solve multi-label task. We also want to fine tune the pretrained BERT embeddings and validate results with other BERT architectures.

References

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. 2004. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771.

Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. 2000. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shuiwang Ji, Lei Tang, Shipeng Yu, and Jieping Ye. 2008. Extracting shared subspace for multi-label classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 381–389.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022a. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022b. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Boris Lauser and Andreas Hotho. 2003. Automatic multi-label subject indexing in a multilingual environment. In *International Conference on Theory and Practice of Digital Libraries*, pages 140–151. Springer.

Jesse Read, Bernhard Pfahringer, and Geoff Holmes. 2008. Multi-label classification using ensembles of pruned sets. In *2008 eighth IEEE international conference on data mining*, pages 995–1000. IEEE.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Rajasekar Venkatesan and Meng Joo Er. 2014. Multi-label classification method based on extreme learning machines. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 619–624. IEEE.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.