Quintilian at SemEval-2023 Task 4: Grouped BERT for Multi-Label classification

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

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

Abstract

In this paper, we initially discuss about the ValueEval (Kiesel et al., 2023) task and the challenges involved in multi-label classification tasks. We tried to approach this task using Natural Language Inference and proposed a Grouped-BERT [1] architecture which leverages commonality between the classes for a multi-label classification task.

1 Introduction

Everyone has a perspective on 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 (Kiesel et al., 2023), is to understand which of these human values form the basis for someone's decision in a textual argument.

ValueEval (Kiesel et al., 2023) 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 options are 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, Mirzakhmedova et al. (2023) also provided each argument with a set of human values in the form of labels that are closely aligned with psychological research. For each argument, these 20 labels provide greater detail about which human values are inferred for making the decision.

2 Background

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 the other classes closer to 0. In a multi-label classification problem, that's not possible as each input 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 PT3 transformation (Boutell et al., 2004) provides better results compared to other transformations. If an input belongs to both the classes A and B, then PT3 transformation generates a new class C, such that input has belonged to only class C (which represents both A & B). Considering that each argument 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, PT4 transformation (Lauser and Hotho, 2003) is preferred as it uses L binary classifiers, with each binary classifier predicting 1/0 for each class.

[1] https://github.com/ajaymopidevi/Grouped-BERT

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.

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 argument in our dataset. For the training dataset, we have observed the LC and LD values in Table 1.

Label Cardinality	3.406
Label Density	0.17

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

Along with the dataset, Mirzakhmedova et al. (2023) provides a few baseline models for comparison. We have the following 2 baseline models: a 20-label classifier network that uses contextual embeddings from a pretrained BERT 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 System Overview

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 this paper, we propose a Grouped-BERT architecture to handle different multi-label classification challenges. As each argument can belong to multiple classes, we tried to group some of the classes that often have the same representation. This grouping can generalize some of the core reasoning behind the arguments, that is common in all its output classes. 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 classes. We combined 4 classes in a group, as our label cardinality, i.e average number of labels per argument is around 3.4

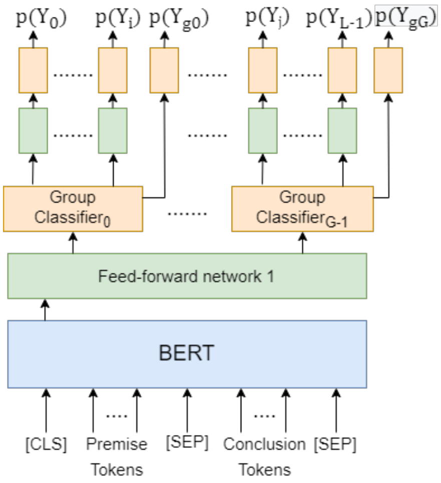

Figure 1: Grouped BERT Architecture, L(=4) represents the number of classes in each group, G(=5) represents the number of groups

The last hidden state of the pre-trained BERT model is initially fed into group classifier, to identify the group of classes it belongs to. Then the output of each group classifier is forwarded to the hidden layers and subsequently the L binary classifiers for each group class, and also to a separate classifier layer to predict whether the final set of labels belongs to that group. We need to create extra labels for these groups for loss calculation, by evaluating its member labels in that group.

4 Experimental Setup

For the input embeddings, we used embeddings generated from pre-trained 'bert-base-uncased' model. We trained our models for 200 epochs with an AdamW optimizer 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. For the loss function, we considered a local loss(sum of CrossEntropyLoss for each group) and a group loss(crossEntropyLoss for the group predictions) and the group is weighted 4 times that of local loss.

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'

The Feed-forward network layer which takes as input the CLS token from the BERT output is a linear layer with 512 output channels. Each of the group classifiers shown in Figure 1 have 64 output channels. The outputs from this hidden layer of 5 group classifiers are then fed into separate networks, each having a single linear layer with 32 output channels. These outputs are finally fed into classifiers, each of which predicts one of the 20 classes.

We trained the model on a laptop with an Nvidia GTX 1070 GPU and Intel 7700HQ CPU.

5 Results

For the main test dataset, our GroupedBERT achieved F_1-score of 0.38, which is less than the baseline BERT model (0.40).

By grouping the classes, we expected that the groups would be independent i.e each argument can have only the values from that group. However, each argument in the training set still belongs to multiple groups, with the label cardinality being 2.54, which is not a much improvement from 3.4. This added more constraints to the original set of labels, rather than identifying the core representation in the grouped labels.

The F_1-scores by our model for the classes Stimulation, Hedonism, Face, Conformity:Interpersonal, Humility is 0. Each of these value categories have the lowest distribution in their respective groups. We suspect that this uneven distribution of labels could be one reason why our model couldn't predict any of these categories correctly.

As we were only able to train our model on 5 year old hardware, we were limited in our ability to train BERT layers and instead had to rely on pre-trained BERT layers with the weights frozen. This could be another contributing factor because of which we could not achieve a better performance with this model.

6 Conclusion

We would like to test our GroupedBERT with a part of the dataset, having even distribution of all the classes and verify the label cardinality stays close to 1. These approaches might provides insights whether GroupedBERT architecture can be used for multi-label classification tasks.

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.

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, 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.

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.

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
2023-01-22-02-01-11	.38	.49	.58	.00	.00	.58	.23	.44	.00	.66	.52	.47	.49	.00	.00	.41	.30	.65	.64	.38	.45
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
New York Times																					
Best per category	.50	.50	.22	.00	.03	.54	.40	.00	.50	.59	.52	.22	.33	1.00	.57	.33	.40	.62	1.00	.03	.46
Best approach	.34	.22	.22	.00	.00	.48	.40	.00	.00	.53	.44	.00	.18	1.00	.20	.12	.29	.55	.33	.00	.36
BERT	.24	.00	.00	.00	.00	.29	.00	.00	.00	.53	.43	.00	.00	.00	.57	.26	.27	.36	.50	.00	.32
1-Baseline	.15	.05	.03	.00	.03	.28	.03	.00	.05	.51	.20	.00	.07	.03	.12	.12	.26	.24	.03	.03	.33

Table 2: Achieved F_1-score of team quintilian per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. Approaches marked with * were not part of the official evaluation. Approaches in gray are shown for comparison: an ensemble using the best participant approach for each individual category; the best participant approach; and the organizer's BERT and 1-Baseline.

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.

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.