Application of text classification approaches for Human Value Detection
NLP Course Project

Domenico Dell'Olio, Giovanni Pio Delvecchio and **Raffaele Disabato**
Master's Degree in Artificial Intelligence, University of Bologna
{ domenico.dellolio, giovanni.delvecchio2, raffaele.disabato}@studio.unibo.it

Abstract

The task of Human Value Detection proposed by Kiesel et Al. through the "Touché23-ValueEval" challenge is focused on extracting human values from arguments. In this work we provide three different models of increasing complexity to tackle this problem: a simple but not trivial baseline consisting of GloVe embeddings + a small GRU RNN, an evolution of the previous using BERT with transfer learning (BERT+LSTM) and a fine-tuned BERT model for text classification. All the models were tested both on a test with the same cultural background of the training set and on one with a different one. The fine-tuned BERT was confirmed to be the best model with 42% and 31% F1-score on the respective test sets, overcoming the BERT+LSTM model by a 5% F1-Score.

1 Introduction

Human Value Detection is a quite recent task, focused on extracting the human values expressed in arguments. A human value can be briefly defined as something we deem as good and right, which can be subjective and guides our decisions and actions. Extracting these values from arguments, that are means of communication where people try to sustain a position or opinion, could be useful in many ways as they provide context to argumentative statements, thus allowing, for example, conditional generation or selection of arguments for a audience with given values, identify common values on controversial topics or help key point analysis (Mirzakhmedova et al., 2023). This task can be included in the class of multi-label text classification problems, as each argument can be labeled with more than one human value. Also, Human Value Detection is at the center of an open challenge - "Touché23-ValueEval" - where a dataset of arguments in form of "Premise" stating the reason of the argument, "Conclusion" giving the context to the argment and "Stance" indicating the relation between "Premise" and "Conclusion" is given to be used for text classification model. The only existing model to this day, dealing with said dataset, are a fine-tuned BERT for classification and a 1-Baseline (model returning always 1 for each label), which reach good results but still having some room for improvements. We propose three different models to approach this task: an "advanced" baseline model obtained with GloVe (Pennington et al., 2014) embeddings followed by 2 GRU layers, an improvement on said baseline with BERT frozen encodings (Devlin et al., 2018) followed by 2 LSTM layers and we re-propose the BERT fine-tuning benchmark with some small modifications. The latter serves also as a more reliable comparation source as we aren't using the full dataset for the task but only the train, validation and Zhihu-validation split, since as per today, not all the labels were published. Differently from the already presented models, our models were inputted the whole argument rather than only the premise and have a different tuned threshold for score-to-label conversion. They were trained on $\approx 77\%$ of the original training split, validated on the rest and then tested on the original validation split and on the Chinese-background validation split to probe their robustness, as value systems change from culture to culture. Evaluation was mainly focused on macro and per-class F1 scores. Our BERT+LSTM managed to improve on the GloVe baseline returning 38% F1 on the first test set (+6%), but is still behind the fine-tuned BERT (42% F1). The scores on the Chinese test set are lower, as expected, due to different reasons (different culture, missing classes, different class distribution) but still follow the trends of the first test and the validation set. Results may improve if the experiments will be reproduced on the full dataset.

2 Background

A human value is defined to be "a belief pertaining to desirable end states or modes of conduct, that transcends specific situations, guides selection or evaluation of behavior, people, and events, and is ordered by importance relative to other values to form a system of value priorities" (Kiesel et al., 2022). Values can also be organized in a hierarchy with different levels of abstraction. The task at hand focuses on the second level of the hierarchy presented in Kiesel et al. (2022) which consists of 20 classes organized in a circular taxonomy. These values can emerge from arguments, which are traditionally defined as "a coherent series of reasons, statements, or facts intended to support or establish a point of view"[1], and generally are subdividable in "premise" (expressing reasons) and "Conclusion" (expressing the context). The main challenge in human value detection from an argument mining point of view, is that values are often not explicit in the argument, but there are clear semantic relations between arguments containing the same values. To this regard, this task can be approached as a multi-label text classification task. The only available works on the dataset propose some common way to deal with this task (SVM, Finetuned BERT, 1-baseline...), mainly focusing on the premise of the argument, in which formally the values of the speaker are expressed (Kiesel et al., 2022). However, there are some cases in which the conclusion helps to disambiguate the premise, therefore we have tested similar models but including also the conclusion and the relation between it and the premise ("Stance") in the inputs. Lastly, we must remark that the relativity of value importance is also due to cultural differences and personal opinions. Therefore tests were performed on a different split containing arguments from another cultural background with respect to those in the training set, in order to probe the robustness of our models.

3 System description

Each example is pre-processed with the appropriate tokenizer (depending on the kind of embedding employed) by padding, truncating, filtering, and lower-casing it. After that, the examples are directly passed to the models to be trained. The models tested are shown in Figure 1. The model with

[1]Definition taken from here.

	examples		unique conclusions	
Train	4176		265	
Val.	1217		67	
Test	1896		157	
Zhihu	100		12	
Length (tokenized)	Premise (mean)	Premise (90° percentile).	Conclusion (mean)	Conclusion (90° percentile).
basic-english	23.74	38	6.66	10
BERT	27.27	42.0	9.9	13.0

Table 1: Statistics on the "Dimensions" of the various splits used.

GloVe (Pennington et al., 2014) embeddings acts as an advanced baseline, since it has relatively few parameters and has a simple architecture usually employed as starting point for text classification. BERT+LSTM uses frozen BERT embeddings (Devlin et al., 2018) to improve the results of the GloVe model and focuses the computation on the output cell state of the Bi-LSTM layers, as ideally they would contain the most relevant information of the whole argument and require less parameters than the ones used for processing a flattened LSTM output. Finally the fine-tuned BERT model was added as it was one of the best performing model in Mirzakhmedova et al. (2023) and could be another possible model to use to tackle this dataset. It has a slightly different architecture (the two last dense layers are not present in the original architecture used in the other work) and has a different threshold for the score-to-labels conversion. In fact, it must be noted that all the models output directly the results of the last layer, without application of an activation function (included in the BCE loss[2]). To perform predictions, logits are passed to a sigmoid function and then a threshold is applied to convert them to hard labels. Finally, all the architectures codes were originally defined, except for the fine-tuned BERT model code which was adapted from a tutorial linked in Sec. 8.

4 Data

The dataset we considered is a part of an extension of the one presented in Kiesel et al. (2022) adapted for the "Touché23-ValueEval" challenge, which is presented in Mirzakhmedova et al. (2023). The full dataset consists of 9324 arguments consisting of three parts - "Premise", "Conclusion" and "Stance" (of the premise w.r.t. the conclusion) - expressing the argument, 20 binary labels for the second level of the taxonomy and 54 for the first one (not used). These examples are subdivided in a "main" dataset

[2]See PyTorch's BCEwithLogitsLoss.

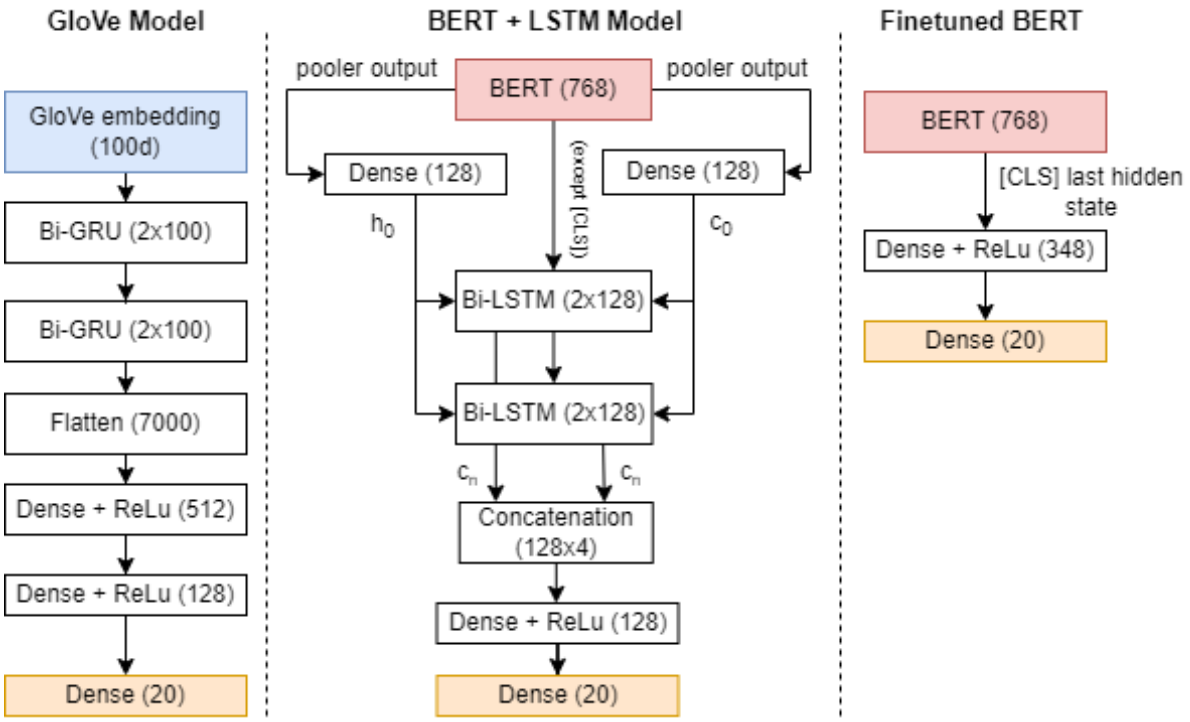

Figure 1: Tested Model architectures. Between the parentheses there are the number of neurons of dense and dense-like layers or the embedding dimension for each token for the recurrent/embedding layers.

(8865 examples) and a "supplementary" one (459 examples). The first one has mainly arguments with American cultural background plus some examples from European (1098 examples) and Indian (399 examples) sources, while the supplementary comprehend three sources: a chinese one (Zhihu), an islamic one (Nahj al-Balagha) and the New York Times. Both parts are provided already split, but since the challenge was still open when this work was being produced, not all the labels were publicly available. Thus, we decided to split the provided training set of the "main" part into our actual training split and the relative validation split for hyper-parameter tuning, and to use the proposed validation split as test set. Finally to probe the robustness of the model, the Chinese part of the "supplementary" dataset was also used as an additional test set. Some statistics on the dimensions of the dataset are in Table 1. The split on the training set was performed so that every example with the same unique conclusion are in the same split, to avoid train-test leakage, as done in the reference paper (Mirzakhmedova et al., 2023). In particular the training set contains the 80% of unique conclusions ($\approx 77\%$ of examples). The distributions of each class is uneven inside each split but they are roughly the same through train, validation and test splits except for "Self direction: Thought" and "Power: resources", which are more present in the validation set. The Chinese split, however, has a quite different distribution of values and misses two classes ("Stimulation" and "Tradition"). Moreover it is important to remark that most of the examples in all the split have at least two values, and brought us also to consider co-occurrencies between classes. About the pre-processing, it is rather light, as each example has the whole argument joined in a single string as ["Premise"]s["Stance"]s["Conclusion"], where s is either a white-space character or the "[SEP]" bert token for model using BERT embeddings.Then, the resulting input is tokenized either with the `basic-english` torch tokenizer or the BERT tokenizer, and then truncated and padded. The maximum token length for the GloVe baseline model is 35 (which is slightly above the sum of the mean number of tokens for both premises and conclusion), while for BERT is 70 (above the sum of the 90^{th} percentiles of lenghts). This is due to the fact that GloVe would return too many zero vectors for both padding and OOV with longer sequences that would interfer with training, while the embedding returned by BERT returns representations containing zero values quite rarely.

5 Experimental setup and results

The GloVe model and the BERT+LSTM model were trained for a maximum of 50 epochs, with Adam optimizer and an early stopping monitoring the validation loss, with a patience of 3 epochs and a min delta of 10^{-4}. Hyperparameters were tuned manually by observing the results and the loss on the validation set:

- Batch size: 32 for both, chosen in {16,32,64}.
- learning rate: 10^{-4} for GloVe and 10^{-3} for BERT+LSTM, chosen in $\{10^{-2}, 10^{-3}, 10^{-4}\}$.
- GRU units for GloVe: 100 chosen in {100,200}.
- LSTM units for BERT+LSTM: 128 chosen in {128,256,512}
- Neurons for linear layers were chosen in {128,256,512}.

Differently, BERT was finetuned for a maximum of 6 epochs, with AdamW optimizer and gradient clipping to 1. The tested hyperparameters were:

- Batch size: 16 between {16,32}.
- Learning rate: 5e-5 chosen between {2e-5, 3e-5, 5e-5}.

We also tuned the threshold for the conversion from the scores returned by the sigmoid to hard labels. Values between [0.2,0.5] were tested, and 0.25 was chosen considering the trade-off between precision and recall. The main reference measure was the macro-F1 score evaluated, but also precision, recall and accuracy were produced, as well as per-class results. Accuracy results are not displayed here, as they are not so informative. The results on different splits are reported in table 2. Other experiments were done but not reported here, as learning only on premises, trying other kind of architectures involving GAT (Veličković et al., 2017), 1D convolutions or other variations on BERT (ELECTRA, FunnelTransformer, BORT...) but they did not return remarkable enhancements.

Model	Macro val. precision	Macro val. recall	Macro Val. F1
GloVe	0.32	0.38	0.31
BERT+LSTM	0.38	0.41	0.35
BERT f.t.	**0.47**	**0.42**	**0.40**
	Macro test precision	Macro test recall	Macro test F1
GloVe	0.41	0.39	0.32
BERT+LSTM	0.38	**0.44**	0.38
BERT f.t.	**0.46**	**0.44**	**0.42**
	Macro zhihu precision	Macro zhihu recall	Test zhihu F1
GloVe	0.19	0.43	0.25
BERT+LSTM	0.20	0.43	0.26
BERT f.t.	**0.28**	**0.44**	**0.31**

Table 2: Results of the runs of the models using seed 10 for reproducibility. "f.t." stands for fine-tuning and "zhihu" refers to the dataset with different cultural background.

6 Discussion

6.1 Results on the test set

As we can see from the tables, the results are quite consistent from the validation to the test set, where the fine-tuned BERT model seems to excel among the three, reaching 40-42% F1, while the BERT+LSTM only improves on the GloVe Baseline by about 5%. We must remark that the BERT+LSTM improvement is achieved by changing the embedding, the RNN layer type (LSTM has more capacity than the GRU, and thus can be more effective) and halving the total trainable parameters required by GloVe ($\approx 3.5M$). These parameters are still a lot less than the one tuned on BERT ($\approx 110M$). We can also observe that the F1-results on the test are slightly higher than the ones on validation because the distribution is more similar to the training set and the classes are better supported. By analyzing per-class scores, we can deduct that:

- The most difficult classes to be predicted (F1-scores<25%) are "Stimulation", "Hedonism", "Face", "Conformity:Interpersonal" and "Humility", all being classes under represented in the training split.
- There are some fluctuation ($\pm \approx 7\%$) on some values of F1 between validation and test set scores, depending on a light skew of the class distribution of the validation split w.r.t. train and test. The affected classes are "Power: resources", "Security: Societal", "Universalism: Tolerance".
- The class "Universalism: Tolerance" (being tolerant to others) is still badly recognized, despite being quite represented in the training dataset (≈ 500 examples). Further analysis on the corresponding examples showed how the most confused classes are some that

have higher representation in the training set, which appear often together and are very near in the values continuum (Kiesel et al., 2022)- Benevolence: caring, Universalism: concern, Self-direction: action i.e. being good to others, being concerned with justice and equality, have freedom of action - or "commonly" opposed to it - Security: personal and Security: societal i.e. having a secure country/environment. This suggests that our models still have some difficulty in distinguishing argument containing these very similar or correlated values from their semantics.

6.2 Results on the different-sourced test set

The results on the chinese test set are less encouraging than the previous ones. They, as expected, drop of $\approx 10\%$ of F1 on all models, with the fine-tuned BERT being still the one reaching the highest score. The improvement registered on the test set between the GloVe model and the BERT+LSTM here is thinned out to a $\approx 1\%$, showing that the model remains still less flexible than the fine-tuned BERT. The general loss of performances is due to different reasons:

- The set is missing two classes (which F1-score is counted as 0 when aggregating).
- Despite having a different distribution, the under-represented classes remain under-represented in this split, proposing 1 to a few example per-class.
- The inherent difference in the cultural background and the themes treated in the set, which draws the examples away from the ones used for training.

7 Conclusion

In this work we presented some possible starting point to address the Human Value Detetection task as posed by the "Touché23-ValueEval" challenge. We proposed first a non-trivial baseline using GloVe embeddings and GRU layers, a variant of this model employing BERT frozen encodings and LSTM layers and a simple SOTA multi-label text classification model obtained by fine-tuning a BERT model. As expected the variant outperformed the first baseline on both validation, test and slightly on a different-cultural-background split, but it wasn't enough to outperform the fine-tuned BERT model, which was one of the first models proposed for this task by the authors of the dataset. However, we must remark that the $\%5$ percent improvement on the BERT+LSTM model came at the cost of more than 109M additional parameters to be tuned/trained. Moreover, all the models still suffer from having under-represented classes and have some difficulty in distinguish more ambiguous values as that of "Tolerance". Probably this could be fixed or improved by using the whole dataset, once the test labels will be available, or trying to inject a different kind of knowledge into the models. This knowledge could be in the form of additional syntactic information (Wang and Li, 2022) or compute an attention score on the labels and then use this information to help classify the argument (Xiao et al., 2019).

8 Links to external resources

- The tutorial considered for finetuning BERT can be found here.
- Learned model weights can be found here.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. 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.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks.

Haitao Wang and Fangbing Li. 2022. A text classification method based on lstm and graph attention network. *Connection Science*, 34(1):2466–2480.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. 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 466–475, Hong Kong, China. Association for Computational Linguistics.