Francis Bacon at SemEval-2023 Task 4: Ensembling BERT and GloVe for Value Identification in Arguments

Kenan Hasanaliyev
Stanford University
kenanhas@stanford.edu
Justin Wu
Stanford Univeristy
justwu@stanford.edu
Kevin Li
Stanford University
kevinli7@stanford.edu
William Huang
Stanford University
willsh@stanford.edu
Saanvi Chawla
Stanford University
saanvic@stanford.edu
Daniel Yang
Stanford University
dy92634@stanford.edu
Michael Nath
Stanford University
mnath@stanford.edu
Shane Mion
Stanford University
smion@stanford.edu
Rohan Sanda
Stanford University
rsanda@stanford.edu
Kiran Bhat
Stanford University
kvbhat@stanford.edu

Abstract

In this paper, we discuss our efforts on SemEval-2023 Task 4, a task to classify the human value categories that an argument draws on. Arguments consist of a premise, conclusion, and the premise's stance on the conclusion. Our team experimented with GloVe embeddings and fine-tuning BERT. We found that an ensembling of BERT and GloVe with RidgeRegression worked the best.

1 Introduction

Identifying the values (e.g., humility and dominance) in an argument is a key part of understanding the psychology of the argument's ideator. Present-day NLP models allow us to more accurately identify arguments for their human values. These models can power better emotional understanding for conversational agents. Based on the values you are exhibiting, a conversational agent would be able to change its response to brighten your mood or de-escalate the situation, resulting in a more engaging experience.

To help advance the NLP effort in value identification, we participated in the SemEval-2023 Task 4 (Kiesel et al., 2023). The SemEval task focuses on classifying the human value categories that a textual argument draws on. The dataset for the task contains arguments consisting of a premise, conclusion, and the premise's stance on the conclusion ("in favor of" or "against"). The target label for each argument is one or more of 20 given human values.

Our team experimented with two approaches for value identification, a baseline model using GloVe word embeddings, and a BERT model that we fine-tuned for multilabel classification of human values. We describe the implementation details of these models in this paper, as well as our experiments and results from these models. We found that ensembling BERT and GloVe acheived a 10% improvement in average F1-score over our baseline GloVe model.

Our full implementation can be found at https://github.com/claserken/MLabTask4SemEval.

2 Background

SemEval-2023 Task 4 consisted of four inputs for each argument: an argument ID, the conclusion, the stance, and the premise. The argument ID is simply a marker for this specific argument (e.g., A01002). The premise is the full argument, such as "we should ban human cloning as it will only cause huge issues when you have a bunch of the same humans running around all acting the same." The stance is a binary choice of whether the premise is supported or not, with this specific one being "in favor of" while others may be "against." The conclusion is drawn from the stance and the premise

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2039–2042
July 13-14, 2023 ©2023 Association for Computational Linguistics

which in this case would be “we should ban human cloning."

All of these inputs would map to an output list of 0s and 1s, like [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0], which represents the 20 human values the argument is indicative of. The datasets were several thousand arguments long and provided in English.

3 System Overview

3.1 GloVe

GloVe (Pennington et al., 2014) is an unsupervised machine learning model that combines some elements of both count based and prediction based methods of learning distributional word representations. It tries to replicate the efficiency of count data, but also utilizes the linear substructures prevalent in log-bilinear prediction-based models. It outperforms most other models on word analogies, word similarities, and named entity recognition tasks, and is fundamentally a log-bilinear linear regression model for unsupervised learning.

Our baseline approach was to use pretrained GloVe word embeddings for the “Conclusion" and “Premise" categories given in the arguments to predict the labels. We started by stripping the text of the words “the", “a", “an", and “of". Furthermore, we removed punctuation, and converted the letters to be all lowercase. We then averaged the embedding vectors for all the words in those categories, obtaining one aggregated word vector for each of “Conclusion" and “Premise". Summing these two vectors together, we obtained the combined word vectors, amounting to one 100-dimensional vector for each argument.

Experimentally, we verified that of the Logistic Regression, Passive Aggressive, Perceptron, Ridge, and SGD Linear Classifiers in sklearn, the Ridge Classifier consistently produced the best results, so we fit the Ridge Classifer with a balanced class weight and calculated the F1 scores with the provided arguments/label training/validation data. We executed this process 20 times, since it specifically takes the entries for one human value every run, and we averaged all the F1 scores (for both the training and validation datasets) to obtain our final results.

3.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a family of masked-language models that uses a deep bidirectional transformer architecture to generate conceptualized word embeddings. It was developed by researchers at Google as a pre-trained language model trained on a large corpora of text. BERT uses the MLM (Masked Language Model) to establish relationships between the words in a sentence, as well as Next Sentence Prediction to solidify the figure out if a relationship exists between different sentences. BERT can be further fine tuned to adapt the model to a specific downstream task such as sentiment analysis.

The base BERT architecture is made up of a stack of transformer encoder layers. After words are tokenized and passed through an embedding layer to obtain vector representations, they are passed through these transformer encoders – which consist of multi-head attention and feed-forward neural networks. Together, these encoders improve the vector representations of the words.

We used the Hugging Face transformers library implementation to fine-tune bert-base-uncased. We added an extra 769-D hidden layer on top of BERT, and an 20-D output layer (D here stands for dimension). The 20-D output layer then goes through a sigmoid function to return probabilities for each of the 20 human values.

We process each argument by feeding the premise through the tokenizer and padding it. The tokenized premise is fed through the BERT model giving us a 768-D premise embedding. This embedding is then concatenated with the binary stance, resulting in a 769-D vector. Finally, this vector goes through the hidden and output layers to obtain the probabilities for each human value.

We used binary cross entropy as our loss function between BERT’s predicted value probabilities and the target probabilities. Since there is a large class imbalance where most values are negative, our model initially predicted negative for each value. As a result, we weighed positive values to have roughly equal weight as the negative values in the loss function.

$$L_{batch} = \frac{|P| + |N|}{2}\left[\frac{\sum_{p \in P} L_p}{|P|} + \frac{\sum_{n \in N} L_n}{|N|}\right] \quad (1)$$

In the equation above, P refers to the positive examples in each batch (i.e., the value is present in the argument). Similarly, N refers to the negative samples in each batch. Dividing by the size of P

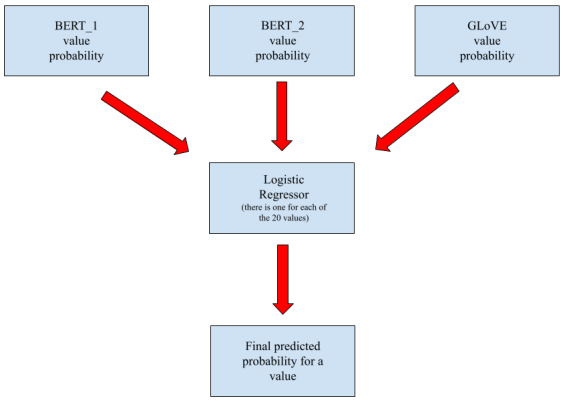

Figure 1: Ensembling architecture diagram. For each of the 20 human values, the BERT and GLoVe predictions are passed through a logistic regressor that outputs the final probability of an argument pertaining to a given value.

and N enabled the model to properly distinguish between positive and negative samples and attain a positive F1 score.

3.3 Ensembling

In an effort to combine our GloVe and BERT models, we ensembled the two. Ensembling was achieving by taking two BERT models with different parameters (batch size, positive samples weight, and epochs trained) and our GloVe + RidgeRegression combination.

For each value, we started the with the outputted predictions from each of the 3 models. These 3 predictions were then fed through a value-specific logisitic regression model and fit to the target labels dataset. Hence, in total, we had 20 different logistic regression models with the idea being each logistic regressor would learn the optimal combination of our BERT and GloVe models. The ensembling architecture is depicted in Figure 1.

4 Results

Table 1 represents the data from the baseline GloVe approach. The rows represent human values, and the columns represent the F1 scores from the training and validation test sets. The F1 score from the training set (0.46) is quite a bit higher than the F1 score from the validation set (0.39), though, suggesting that this approach can be made more complex to decrease the gap. This led to our attempt at our ensembling approach.

Table 1: GloVe + RidgeRegression

Human Value	**Train F1**	**Val F1**
Self-direction: thought	0.56	0.40
Self-direction: action	0.56	0.46
Stimulation	0.19	0.26
Hedonism	0.22	0.20
Achievement	0.61	0.57
Power: dominance	0.37	0.27
Power: resources	0.52	0.43
Face	0.28	0.21
Security: personal	0.68	0.67
Security: societal	0.67	0.55
Tradition	0.47	0.40
Conformity: rules	0.49	0.46
Conformity: interpersonal	0.27	0.12
Humility	0.32	0.14
Benevolence: caring	0.47	0.56
Benevolence: dependability	0.35	0.27
Universalism: concern	0.65	0.57
Universalism: nature	0.54	0.61
Universalism: tolerance	0.40	0.25
Universalism: objectivity	0.47	0.44
Average	0.46	0.39

Table 2 represents the data from the ensembling approach. The rows and columns represent the same categories as before, and here, we can see that both the F1 score from the training set (0.76) and the F1 score from the validation set (0.45) are greater than their corresponding values in Table 1. In particular, the validation F1 score of ensembling is roughly 10% greater than that of baseline GloVe.

For comparison purposes, the 1-baseline provided for this task averages out to a 0.26 F1 score across all categories, with a modified BERT baseline taking on a value of 0.42. Both of our training F1 scores outperformed these benchmarks, whereas the 0.39 validation F1 score from the GloVe approach alone performed worse than standard BERT. However, after ensembling, the validation F1 score increased to 0.45, demonstrating better performance than either of its individual approaches.

Table 2: Ensembling

Human Value	**Train F1**	**Val F1**
Self-direction: thought	0.84	0.51
Self-direction: action	0.87	0.53
Stimulation	0.61	0.32
Hedonism	0.69	0.36
Achievement	0.86	0.64
Power: dominance	0.68	0.34
Power: resources	0.89	0.45
Face	0.62	0.27
Security: personal	0.88	0.73
Security: societal	0.88	0.63
Tradition	0.78	0.43
Conformity: rules	0.80	0.52
Conformity: interpersonal	0.73	0.17
Humility	0.67	0.12
Benevolence: caring	0.66	0.58
Benevolence: dependability	0.66	0.28
Universalism: concern	0.81	0.64
Universalism: nature	0.91	0.68
Universalism: tolerance	0.70	0.26
Universalism: objectivity	0.76	0.45
Average	0.76	0.45

5 Conclusion

Our research began by utilizing GloVe embeddings as a baseline approach. However, we later modified our procedure by using BERT, a powerful language model. Due to class imbalances, we had to reweigh the examples in the loss function. To further improve the performance of our model, we employed an ensemble method that combined the outputs of both GloVe and BERT, and observed a 10% increase in F1 score as compared to the GLoVe baseline.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D 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.