Moral Embeddings

Performance, Generalisability and Transferability

Author: Dragos-Paul Vecerdea - CSE 3000 Research Project - Supervised by E. Liscio and P. Murukannaiah

Introduction

Personal values are the abstract motivations that drive our opinions and actions. Using state-of-the-art NLP methods, we design a classifier to study their expression in text.

Moral Foundations Theory [1] (MFT) proposes five "irreducible basic elements" of morality, that we can frame our study in: *care/harm, authority/subversion, fairness/cheating, loyalty/betrayal, purity/degradation.*

Embeddings convert word and sentences to meaningful vectors and they are an important step in a text classifier's pipeline. They can be domain-adapted to improve the model's performance.

Research Goal: Train embeddings to learn moral foundations and assess our method by answering three research questions:

1) Does our fine-tuning method increase the moral classifier's **performance**.
2) Do fine-tuned embeddings **generalise** across domains of discourse.
3) Are fine-tuned embeddings **transferable**.

Motivation: no prior moral classifiers focus on fine-tuning state-of-the-art embeddings (Sentence-BERT) to improve the model's performance. Moreover, after training, embeddings' utility is not limited to classification task: Semantic Textual Similarity, clustering.

Methods

1. Moral Foundations Twitter Corpus

7 datasets

ALM | BLM | Elections | Davidson | Sandy | MeToo | Baltimore

35 000 annotated tweets

3-8 annotators, moral values or non-moral

Preprocessing

- Lemmatise
- Spell correction
- Remove urls, mentions
- Remove symbols
- Remove punctuation
- Emojis -> Words

Labelling

Majority vote
No consensus -> non-moral
11 classes

Data Balancing

From train data, remove from non-moral class until it equals second most often class

2.Sentence-BERT

Sentence embeddings with state-of-the-art results on Semantic Textual Similarity.

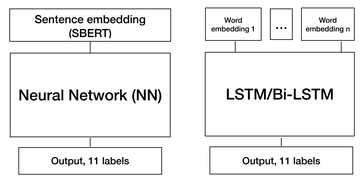
Triplet loss is used to fine-tune pre-trained embeddings

3. Classifiers

We define our problem as a **multi-label** classification problem.

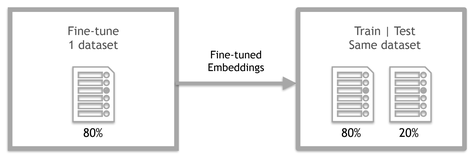
Two baseline models: LSTM and BiLSTM with GloVe word embeddings.

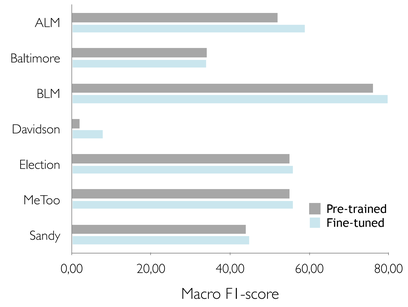
Third model: **Sentence-BERT** embeddings (pre-trained/fine-tuned) with a simple Neural Network on top.

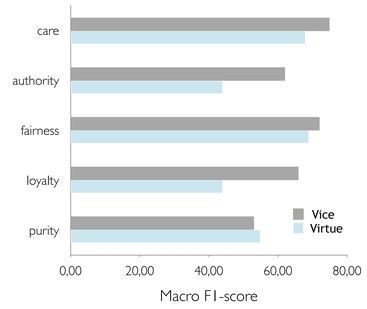


Performance

1) Does fine-tuning Sentence-BERT embeddings with triplet loss increase the moral classifier's **performance**.



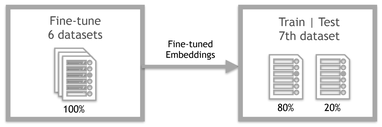




The chart on the left captures how fine-tuning SBERT improves the Macro F1-Scores for the moral classifiers. On the right, we show how embeddings trained on entire MFTC recognise each moral value.

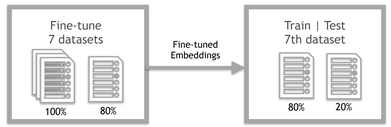
Generalisability

2) Do fine-tuned embeddings **generalise** across domains of discourse.

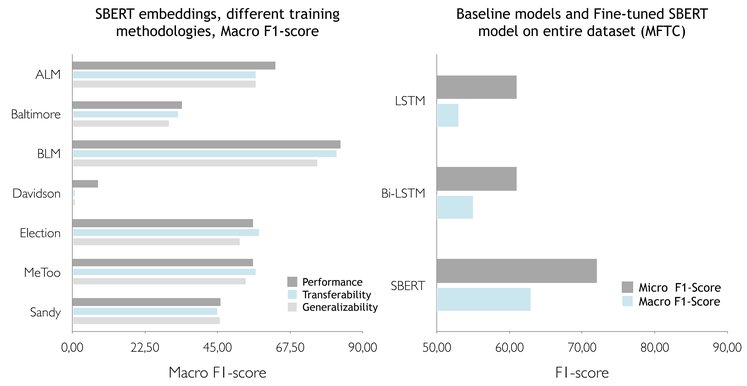


Transferability

3) Are fine-tuned embeddings **transferable**.



Results



Discussion

Understanding the underlying moral drives of people's choices could enable better policy-making and design of virtue-aligned AI. For the moral classification task, we proposed a method to fine-tune state-of-the-art embeddings. The resulting classifier achieves 72% Micro F1-score on the MFTC dataset.

Future work

For a complete understanding of moral embedding's transferability, MFTC should be extended. As MFT annotating is labour intensive, we recommend experimenting with semi-supervised annotating methods [2].

To better explain our method's success, it should be investigated if *semantically similar text expresses similar moral values.*

References

1. Graham, F. (2013). Moral Foundations Theory. *Advances in Experimental Social Psychology., 47, 55-130.*
2. Settles, B. (2011). Closing The Loop: Fast, Interactive Semi-Supervised Annotation with Queries on Features and Instances.

D.P.Vecerdea@student.tudelft.nl