Interpretability of state-of-the-art NLP models for moral values prediction

Ionut Constantinescu - *supervised by Enrico Liscio and Pradeep Murukannaiah*

TU Delft

Background

Understanding personal values is essential for the creation of value-aligned artificial agents that can operate among us.

Tweets are a natural environment where people express their thoughts.

Why did the model predict Y given input X?

Input → Black box → Output

Why do we need interpretability?

- Better models and less bias
- More accountable ML systems
- More trust in ML systems

1

Methodology

MFTC Dataset

7 corpuses

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

35k annotated tweets

3-8 annotators, moral values or non-moral label

5 moral foundations

Care-Harm | Fairness-Cheating | Loyalty-Betrayal | Authority-Subversion | Purity-Degradation

Model training

LSTM	BERT	FastText
Recurrent neural network arhitecture	Bidirectional Transformers arhitecture	Faster text classification with similar results

Goal: Compare the three models based on their interpretability

Interpretability analysis

Experiment 1: Performance
Q: How accurate/reliable are the predictions?

Experiment 2: Input data
Q: What kind of data does the system learn from?

Experiment 3: Embeddings
Q: How does the model extract features from the data?

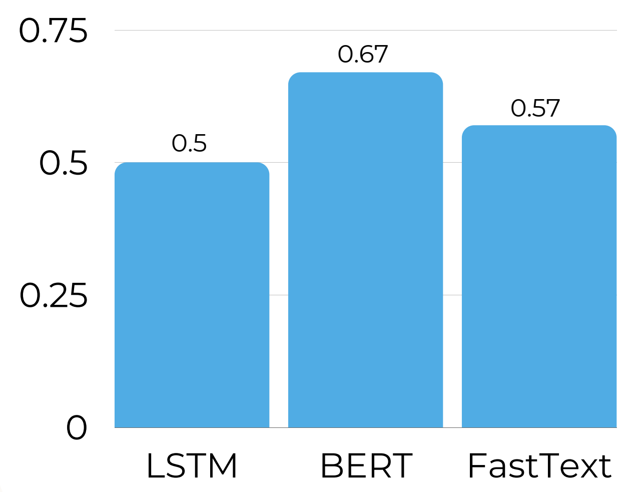
Experiment 4: Feature Attribution
Q: What instance feature leads to the system's prediction

Experiment 5: Counterfactuals
Q: What would the system predict if this instance feature changes to ... ?

2

Experiments and Results

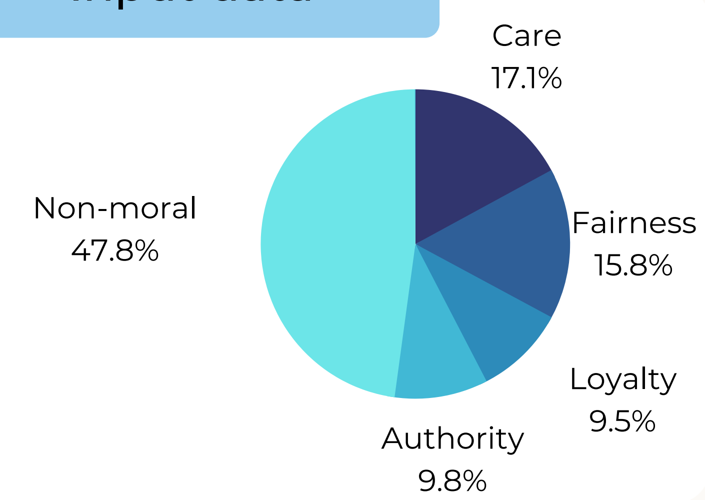
Performance



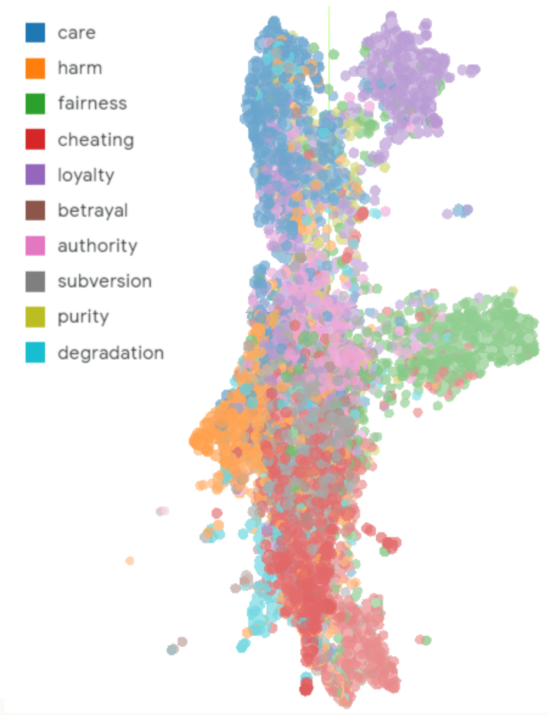
BERT achieves the best performance

Input data

The unbalanced data affects models' behaviour

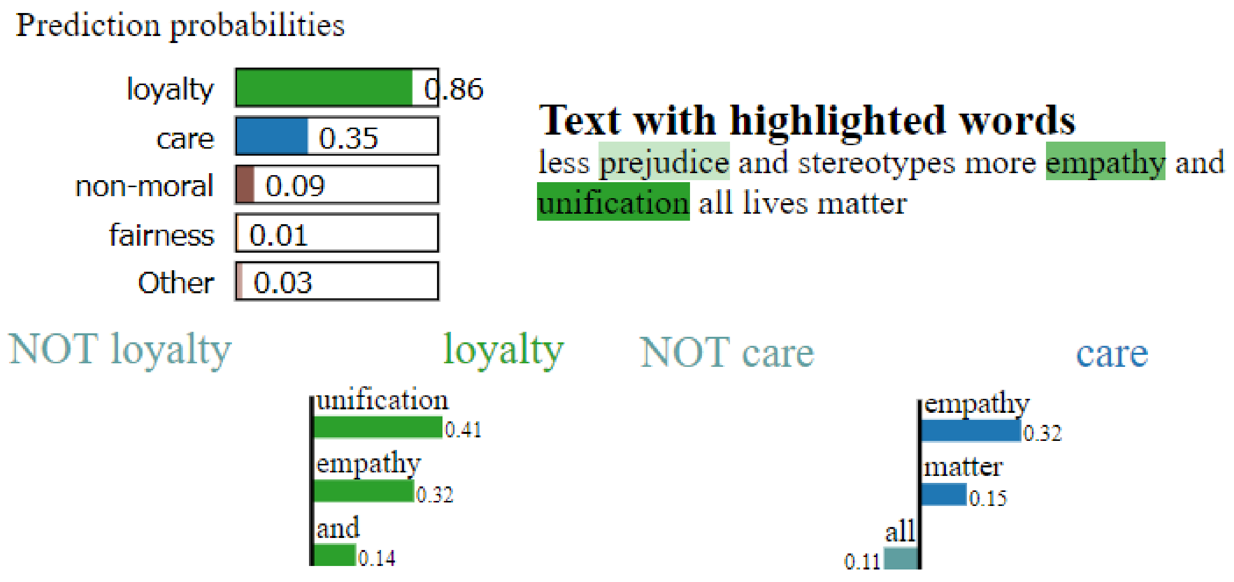


Embeddings



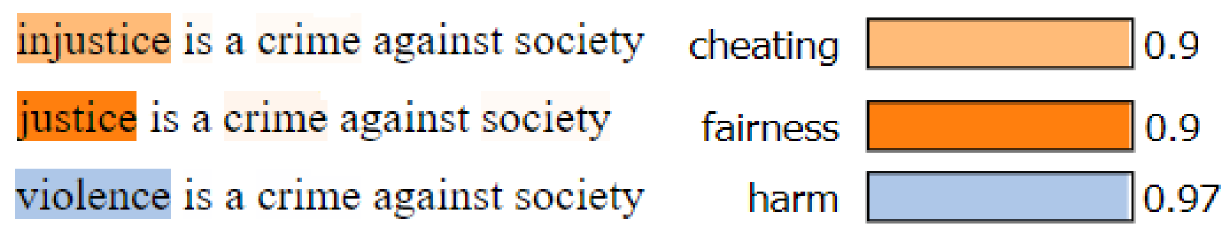
Contextualised embeddings are better at distinguishing between moral classes

Feature Attribution



- *Frequent and meaningful words have a high impact on the predictions*
- *BERT is better at differentiating words by context and noticing semantic particularities*

Counterfactuals



- *A single word can change models' predictions*
- *The behaviour is unexpected when having more than two labels*

3