Making an ethical judgement is not a trivial task, a thorough moral textbook should be tailored to teaching the machine how to differentiate between right and wrong.

'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models

Areej Alhassan[1], Jinkai Zhang[2] and Viktor Schlegel[2]

INTRODUCTION

- (AITA?) is a subreddit on Reddit dedicated to passing moral judgement on everyday conflicts.
- A person can describe a situation that led to a conflict.
- Commenters cast one of four different votes.

DATASET

Our dataset contains 175K posts.

2 steps for collecting **Reddit** posts:

1. Pushshift API. **2.** PRAW

Title	AITA for being upset with my family?
Body	I am the middle child in a more than dysfunctional family. My relationship with my mum in particular has always been strained. …. etc
Verdict	Not the A******

Verdict	Label
YTA (You're the A******)	1
ESH (Everyone sucks here)	
NTA (Not the A******)	0
NAH (No A****** here)	

BALANCING THE DATASET

Name of Dataset	Word Count	Verdict
Dataset1 (Imbalanced)	>10 Words <1993 Words	1: 24,000 0: 86,000
Subset2 (Balanced)	>316 Words <512 Words	1: 24,000 0: 24,000
Subset3 (Balanced)	>512 Words <1994 Words	1: 24,000 0: 24,000

EXPERIMENT

Model	Experiment1	Experiment2
BERT	Dataset1	Subset2
RoBERTa	Dataset1	Subset2
$RoBERTa_{Large}$	Subset2	Subset2
ALBERT	Dataset1	Subset2
Longformer	Dataset1	Subset3

EXPERIMENT RESULTS

Dataset	Model	Validation Accuracy	MCC
Dataset1	BERT	0.78	0.091
	RoBERTA	0.78	0.098
Subset2	RoBERTA	**0.81**	**0.644**
Subset3	Longformer	**0.88**	**0.77**

DISCUSSION

Dataset1:

- models were only capable of predicting the majority class.
- data imbalance is a challenge.

Subset2:

- RoBERTa-large did not outperform the base version.

Subset3:

- Longformer showed its ability to learn dependencies contained in the long sequences.

FUTURE WORK

- Boost the minority class performance and utilise another balancing technique(e.g. SMOTE)
- Consider multilabel classification.
- External ethical knowledge can be used to improve the rationality of verdicts.

13th Edition of the Language Resources and Evaluation Conference (LREC 2022)

King Saud University

The University of Manchester

Take a picture to download the full paper