Can Machines Detect if you're a Jerk?

Parvathy Sarat
Georgia Institute of Technology
psarat3@gatech.edu

Prathik Kaundinya
Georgia Institute of Technology
prathik.k@gatech.edu

Rohit Mujumdar
Georgia Institute of Technology
rohitmujumdar@gatech.edu

Sahith Dambekodi
Georgia Institute of Technology
sdambekodi3@gatech.edu

Abstract

Humans have a hard time dealing with moral dilemmas as they are often affected by inbuilt bias. Often, they anonymously seek external validation and judgment on the decisions they made in certain tricky situations. In the world of online communities and social media, r/AmItheAsshole Subreddit is a popular platform where users (called Redditors) post moral situations from their lives and other Redditors act as a jury and vote to decide if the writer took an unethical decision. In recent years, Deep Learning methods have also been scrutinized due to the possible biases in models. These biases stem mainly from the data that the model is trained on and this data is influenced by the context of the data collection (for example, criminal profiling in America displaying racial bias). In this study, we attempt to understand how a machine performs in a task that is entirely subjective but is possibly objective since the label is decided after voting from thousands of different people. We use language models like BERT to assess if we can replicate the sentiments shared by Redditors and classify the Redditor's original post according to the verdict that was declared by rest of the Redditors. We then analyze the results to note whether BERT was successfully biased or trained in morally grey situations. We also use transfer learning to explore whether these learned models can be used for other similar tasks.

1. Introduction and Background

Chief Justice John Roberts once said "Can you foresee a day when smart machines, driven with artificial intelligence, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?". Indeed, with artificial intelligence (AI) concepts being applied to cutting-edge work in most fields today, it is reasonable to discuss the potential applications of AI toward judging the moral soundness of acts, and helping address ethical dilemmas effectively. The argument for allowing AI to evaluate moral decisions largely hinges on the perception that implicit bias (i.e., not arising from training data) in decisions made by machine learning (ML) models is significantly less than those made by humans.

In this study, we analyze an application of Natural Language Processing (NLP) where the answer is more subjective than objective which falls in the realm of moral decisions and the resolution of ethical dilemmas. The r/AmItheAsshole Subreddit is tailor-made for this analysis, as the entire premise consists of a user posting a situation (typically involving an ethical dilemma) from his/her/their life. Other members of the Subreddit then provide their comments and judgement on whether the choices made by the user were Morally Acceptable (referred to as MA) or Morally Unacceptable (referred to as MU). The overall verdict is decided by a simple majority vote. These social situations cover a wide variety of situations which can range from the harmless ("*Was I wrong to not allow my daughter to play with my favourite toy?*") to the far more morally complex and serious ("*Should I forgive my parents for lying about my adoption status 10 years ago*"?). Note that an important part of analysis will have to include the fact that these posts are always going to be biased towards the user who writes the post as they are usually trying to get validation for the correctness of their actions rather than the opposite. This places a limit on the correctness of the input as it won't always have the full unbiased context available.

In summary, the problem we seek to analyze is two-fold, a) How in line is the bias from the models with the biases of the users on the Subreddit?, and b) Can the model correctly filter the bias from the user's post to provide an "objective" judgement on that person's actions.

We apply BERT [2] (a state-of-the-art transformer-based deep learning model) in order to effectively encode information from these posts, and build a binary classifier that attempts to predict the verdict delivered by the users of the Subreddit. In addition, we also utilize similar BERT-based models (ALBERT, RoBERTa) to evaluate the differences in performance with the adoption of these variants. We make comparisons between these techniques and attempt to draw insights on prediction characteristics of our models. We choose BERT primarily because it is the most widely used language model in natural language research and has also been shown to work well on similar binary sentiment analysis tasks like SST-2. BERT has also been trained on BookCorpus and English Wikipedia which are not similar to the dataset from r/AmItheAsshole. We also wanted to explore whether this learned knowledge can transfer over to different tasks in a similar domain of sentiment classification like SST-2[14].

The task is relevant to the domain of bias and fairness in AI, since one of the biggest issues that NLP models face is the implicit bias learned from training data. Addressing this task successfully would reiterate the fact that language models do learn and harbor the same biases as the humans who are responsible for the creation the content the models are trained on. This reinforces the need to address the pressing issue of bias in NLP by encouraging work in developing bias mitigation and dataset debiasing algorithms. It also throws light upon the need for explainability in these models and transparency, especially if they are deployed in sensitive real world situations.

2. Related Work

Our classification task can be loosely translated to being able to assess if we can replicate the beliefs, notions and biases a majority of Redditors in our training data share. Since we use contextual word representations based on transformers (as opposed to word embeddings) to encode our input instances, it becomes imperative to survey previous work done around bias in contextualized word representations. Tan and Celis [15] analyze the extent to which state-of-the-art models such as BERT and GPT-2, encode biases with respect to gender, race, and intersectional identities. While the novelty of their approach lies in evaluating bias effects at the contextual word level, as opposed to at the sentence level, they find that racial bias is strongly encoded incontextual word models, and observe that bias effects for intersectional minorities are exacerbated beyond their constituent minority identities. This finding gives us two perspectives to look at our problem statement. If we are successful in our task, it would mean that the corpus we trained our dataset on (and hence, in turn, the Redditors who were responsible for the creation of that content) harbor biases similar to contemporary contextual word embeddings and that model can also mimic these biases very easily. In addition, that would also imply that while our classifier would project similar views as the Subreddit members, it would not possess any objectivity when providing a judgement if the poster was unethical.

Another related work is by Jentzsch et al. [4], in which the authors demonstrate that machine learning can learn not only stereotypical biases but also answers to ethical choices from textual data that reflect everyday human culture. Their pipeline first uses Word Embedding Association Tests (WEAT) to extract verbs denoting actions. Then, a method called Moral Choice Machine (which contains templates of moral questions, such as “Should I kill people?” with answer templates of “Yes/no”) is used to inspect presence of human biases in text. The model’s bias score is the difference between the model’s positive and negative response scores averaged for all QA templates with that choice. The correlation of WEAT values and moral bias is examined to compute final results, which indicate that text corpora contain recoverable and accurate imprints of human social, ethical and moral choices. However, this method relies solely on unigram verbs (signifying an action) in a sentence, which contain no context of the larger picture being narrated in the sentence or text. Besides, there is no component component in the pipeline that learns or detects biases using data and applying deep learning methods on them. Schramowski et al.[12] built and improved upon this work to show that an advanced semantic representation of text, such as BERT allows one to get better insights of moral and ethical values implicitly represented in text. They not only focus on atomic actions (unigram verbs) but also move to more complex actions with surrounding contextual information and show that BERT has a more accurate reflection of moral values.

Kurita et al. [5] have investigated creating a template based method to measure the bias in BERT word embeddings. They give BERT a log probability bias score specifically in gender bias situations showing that certain attributes are overly common with one gender rather than the other.

Siqi Liu.[7] attempted sentiment analysis of Yelp Reviews and compared the performance across different machine learning and deep learning models. They found that simpler models such as logistic regression performed better than larger models including BERT. This appears to be in line with previous work from Andreea Salinca. [11] as well as the related work on the AITA dataset [9, 10] where a logistic regression model performed marginally better than BERT.

3. Dataset and Prior Work

We use the AITA dataset [10] for our training and testing purposes. This dataset consists of content scraped from the Subreddit r/AmITheAsshole from 2012 to January 1, 2020. Although the verdicts are handed down in four degrees of increasing severity, we combine them into two distinct categories (MU and MA). There are a total of 80k posts.

These posts consist of both the title and the body of the post along with the label.

Previous work on this dataset has been limited to using simpler machine learning algorithms and also using base BERT for the classification task. One of the previous studies [10] attempted to create a baseline for this problem with a basic classifier of logistic regression using 1-gram frequencies of post titles and bodies as features. With 5-fold cross-validation, this classifier performed above-chance with a modest prediction accuracy of 0.62.

The other work from Andrei Mircea[9] used a BERT language model to train a classifier. By training on ≈30K posts, this effort reported an accuracy of 0.61, surprisingly lower than logistic regression which relies heavily on word count and does not account for the sequences amongst words. However, in this method, there were computational limitations which affected the performance, including truncating representation to 512 tokens, restricting the task to binary classification, among others. Moreover, the authors only used the smallest version of BERT (and not more recent versions like ALBERT or RoBERTa). There was also no implementation a transfer learning approach from models trained on other similar sentiment analysis datasets.

4. Models

We trained multiple different variants of BERT to capture a wider range of metrics. This allowed us to directly compare results and judge how they vary across different training methodologies of the same model. ALBERT [6] is more computationally efficient than BERT while RoBERTa [8] has better performance on public datasets like the GLUE benchmark. We train all three of these models across different model hyper parameters for better output diversity. We then used the best performing model for our analysis of the dataset. Specifically what words/concepts occur most commonly with MA or MU labels.
We also tried a transfer learning approach to see whether the learned knowledge in this dataset would generalize to other similar binary classification problems. Specifically we focused on the sentiment analysis dataset SST-2. We applied the best performing ALBERT model since it was the most computationally efficient and thus the quickest to get results without too much performance sacrifice.

5. Limitations and challenges encountered

We encountered two issues while working on this project.

1. One of the issues we encountered was in the distribution of samples in the dataset. While it was ensured that the data was divided into train, validation and test sets after performing stratification (i.e., making sure that the class distribution in all three sets was roughly similar), the dataset was inherently slightly imbalanced. Two potential solutions were explored, but not adopted to address this. As a first, we attempted to undersample the majority class (MA-Morally Acceptable) to a similar level as the number of occurrences of MU (Morally Unacceptable) verdicts in our dataset, so that the ratio $\frac{n(MA)}{n(MU)} \approx 1$. However, on training models in this setting, there was a notable decrease in accuracy, suggesting that we were losing important information. Another possibility that was explored was the application of an oversampling method to the MU set, such as SMOTE [1]. However, synthetic data generation for text samples with SMOTE is problematic and difficult to comprehend, as it is impossible to relate the artificially synthesized data to features in the input space (text). As such, we could not verify the fidelity of the additional samples generated. In the end, we accepted the class imbalance, and account for the possibility that our results may be slightly biased toward the majority class.

2. Another problem we encountered was with the amount of compute resources. While we used a GPU for our experiments, many of the pre-trained models gave us memory issues when we tried to increase the batch size of our training data and/or maximum length (in words) of each instance of training input. This prevented us from experimenting with higher batch sizes and using a larger amount of training data per input instance. We were able to achieve a maximum batch size of 64 samples for each of the three models implemented.

6. Experiments and Results

6.1. Experiments

6.1.1 Data Collection and Preprocessing

We chose the data set collected and cleaned by [10] which has posts dating from the first post in 2012 to January 1, 2020. After cleaning and choosing only those posts with a score of 3 or more (score is the number of upvotes minus the number of downvotes), the dataset had ≈ 63k posts. Each data point contains an official id code, timestamp, post title, post text, verdict, score, and comment count. Since the data was already cleaned, we didn't need to do any preprocessing on it. The only preprocessing required was the tokenization and encoding needed to convert the input data into an appropriate format so that each sentence can be sent to the pretrained models to obtain the corresponding embeddings. The entire dataset was then partitioned into an $70\% - 20\% - 10\%$ split to form the train, validation and test sets, respectively.

6.1.2 Primary Experiments

We implemented our code in PyTorch, and utilized an NVIDIA Tesla V100 GPU to train the classifiers. The two main parts of this project were:

1. *Training and evaluating pre-trained language models on the AITA dataset to compare their performance :* We experimented with the following models, using their corresponding tokenizers in preprocessing -

   BERT : Base and large (uncased) models
   RoBERTa : Base model
   ALBERT : Base and large models

   For BERT and RoBERTA, the posts had to be truncated to 512 tokens by design, while the maximum length was set to 64 for ALBERT. We have used the process of fine-tuning, where, for each of the pre-trained models we added one fully connected layer which learns parameters and then used logits from softmax to get probabilities. In our fully connected layer architecture, we have a linear layer, with a ReLU and a dropout layer added for regularization.

2. *Transfer Learning:* We also applied our trained model onto other sentiment analysis datasets specifically SST-2 which we accessed through the GLUE benchmark.

The loss function and the optimizer used were cross-entropy and the Adam optimizer respectively. For each of the models, we experimented with the following hyperparameters: batch size and learning rate. We chose to experiment with our hyperparameters over the following ranges - batch size of 32 and 64, learning rates of 2e-5 and 5e-5. The batch size range was chosen with regards to the effect it has on training dynamics. While it is known that the effects of hyperparameters as behavior often varies from dataset to dataset and model to model. The works [3] and [13] evaluate the effect of batch size in terms of the generalization gap and find that higher batch sizes leads to lower asymptotic test accuracy. The learning rates were chosen to be of the modest order of $e-5$.

6.2. Results

For the classification task, we used model accuracy and F1-score on the held-out set as metrics for comparing performance, taking into consideration the imbalance in classes. All our models surpassed the existing accuracy baseline of 61%, with ALBERT obtaining highest accuracy of 73.55% (with batch size 64 and 4 epochs). Our best model considering both accuracy and F1-score was BERT with a batch size of 64 and learning rate of 2e-5, achieving a test accuracy of 63.94%. The results of all the models implemented in this study are outlined in Table 1. In the remainder of this discussion, we refer to the specific case of the BERT model with a batch size of 64 and learning rate of $2e-5$. For this specific model of BERT we used early stopping according to 1 as the model was overfitting beyond $\approx$120k iterations.

Model	Batch Size	LR	Val Acc	Test Acc
BERT	32	2e-5	71.81	63.89
BERT	32	5e-5	71.39	64.29
BERT	**64**	**2e-5**	**71.85**	**63.94**
BERT	64	5e-5	71.44	64.63
RoBERTa	32	2e-5	72.73	64.49
RoBERTa	32	5e-5	72.63	73.55
RoBERTa	64	2e-5	72.87	66.12
RoBERTa	64	5e-5	72.63	73.55
ALBERT	32	2e-5	72.63	72.64
ALBERT	32	5e-5	72.63	70.72
ALBERT	64	2e-5	72.65	73.55
ALBERT	64	5e-5	72.37	73.55

Table 1. **Model Performance on Train & Test Sets**

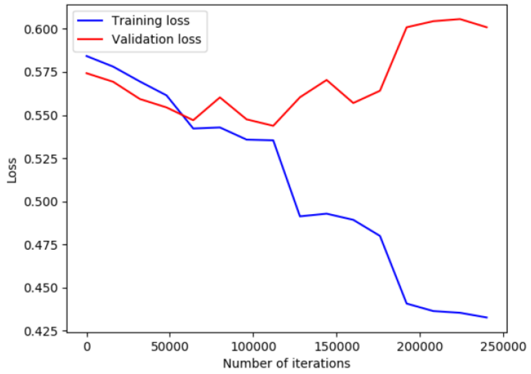

Figure 1. **Learning Curve for BERT Model (64 Batch Size, 2e-5 Learning Rate)**

As elaborated earlier, the imbalanced dataset led to a biased model. The plot for the ROC curve for the best model, as shown in Fig.2 illustrates the effect of class imbalance. The confusion matrices for our best and worst models (by F1 scores) attached in Fig.3 clearly display the biased predictions.

For the transfer learning approach the results were not very favourable. The validation accuracy on SST-2 without any fine-tuning was 47.31%. After fine-tuning, this accuracy goes 51.38% which is an improvement but still not showing good performance. This in contrast to the pre-trained ALBERT model on SST-2 which shows 89.3% validation accuracy. This shows that while this dataset is useful to analyze for bias, it does not transfer well to other sentiment analysis type problems very well. Datasets that more closely mimic the task of the r/AITA data may give better results but we were unable to find a close match which already had public results.

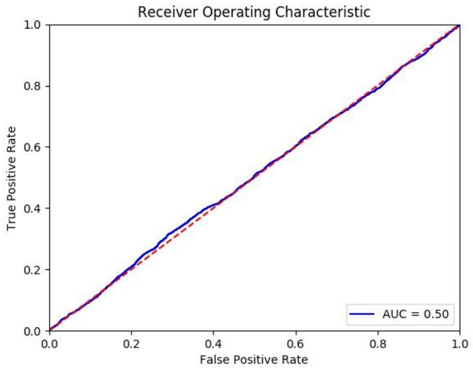

Figure 2. **ROC curve for the best model : Class imbalance resulted in high accuracy but low AUC & F1-scores**

Qualitative Analysis : Probing into the classification results from our best classifier, we see an overwhelmingly large number of words common to both MU and MA classes indicating the learned differentiators must be more nuanced, and a majority of the posts on r/AITA revolve around friends, family, work and home. Posts misclassified with high probabilities by the model tend to follow the same overall behavior.

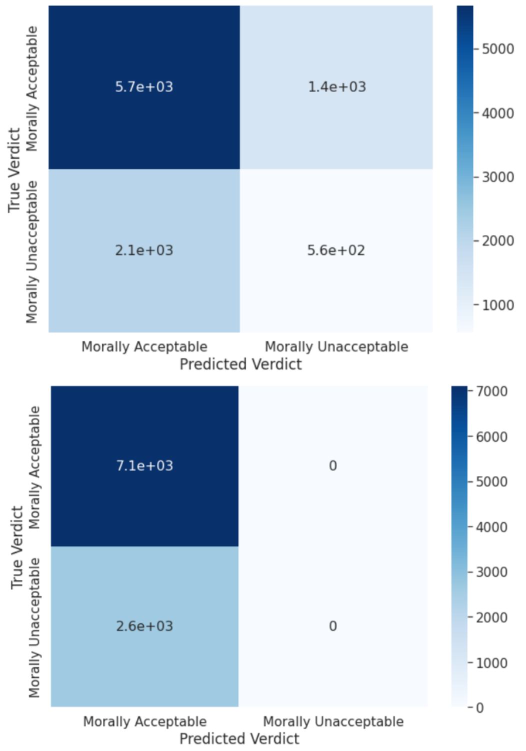

Figure 3. **Confusion matrix for our best (top) and worst (bottom) models by F1 score**

We also see cases where our model is alternately more forgiving and more harsh in its verdict than the Redditors. While the audience originally voted the post *AITA for timing my coworkers smoke breaks?* as MA, our classifier strongly classified it as MU. On the other hand, to *AITA for waking my husband up early to help with the baby?*, the response was the reverse. It might also be that the pre-trained model with fine-tuned weights is on murky ground as it learns a new ground truth from the dataset.

We also assessed what words occurred commonly in posts that were classified correctly to see if there are patterns that our models and the Redditors universally agree on. Relying merely on frequency count can be misleading, even if we remove stopwords. Hence, we use tf-idf (term frequency-inverse document frequency) instead of word frequencies to balance out the term frequency (how often the word appears in the document) with its inverse document frequency (how often the term appears across all documents in the data set), where we treat each post as a document. Words with higher tf-idf in a post would denote greater relevance or influence and would hence be loosely indicative of the topics of posts that drive Redditors judgements. Instead of using the words with the highest tf-idf scores per document (we can call them the 'top words'), we calculate the top words across the corpus for both cases - corpus formed by documents classified correctly as MU and the separate corpus formed by documents classified correctly as MA , as shown in Table 2. We calculate the average tf-idf score of all words across all documents, i.e. the average per column of a tf-idf matrix. It is important to first filter out the words with relatively low scores (smaller than the provided threshold). This is because common stop words, such as 'a' or 'the', while having low tf-idf scores within each document, are so frequent that when averaged over all documents they would otherwise easily dominate all other terms.

We see that the terms probably most relevant to both corpuses are extremely similar - they revolve around the topics of family, relationships, home and work. It is likely that given the highly personal nature of the topics, Redditors perhaps face the most moral conflicts in these areas. It could be inferred from this that when it comes these topics, perhaps, both our model and the Redditors might dole out the same verdict. This is indicative that the model might have learned the same beliefs/notions around these topics as that of an average Redditor. We also notice that the posts related to female topics like mom or wife tend to be not only be more common but also more divisive. However, wife is much higher in the True Positive table indicating that for BERT, marriage posts tend to correctly be placed in the MA class. The demographic of Reddit users skew towards males in the 19-29 age range which indicates that males are more morally unsure about situations involving females.

Feature	tf-idf
mom	0.013625
friend	0.011677
sister	0.011442
dad	0.010284
family	0.010106
friends	0.008785
dog	0.008670
wife	0.008495
kids	0.008472
car	0.008461
money	0.008331
brother	0.008101
parents	0.007982
husband	0.007741
work	0.007703
said	0.007672
just	0.007244
house	0.007058
mother	0.006777

Feature	tf-idf
wife	0.013198
sister	0.013105
mom	0.013059
friend	0.011604
friends	0.011216
money	0.011021
dog	0.010586
just	0.010552
boyfriend	0.010501
car	0.010403
brother	0.009534
job	0.009198
said	0.009198
school	0.008901
parents	0.008834
kids	0.008831
girlfriend	0.008721
dad	0.008600
house	0.008529

Table 2. Left: **Top tf-idf words for correctly classified MA posts** Right: **Top tf-idf words for correctly classified MU posts**

7. Conclusion

We trained a variety of BERT models on the r/AITA dataset and all of the models performed better than the previous baseline in terms of test accuracy. This means we managed to make all our models almost as biased as the average Redditor. Our best model was BERT in we managed to correctly predict 63.94% which is better than the previous baseline of 61%. This BERT model also achieved the highest F1 score which is why we picked it for our analysis. We highlighted certain cases where BERT was seemingly irrationally much more harsh than Redditors. Since BERT was originally trained on Wikipedia text we suspect that there is some implicit information from that dataset that is informing the outputs in a less biased way in some form. We noted in our analysis that the posts are highly female-centric in a male dominated website and the BERT model picks up on this. This could indicate the model learning misogyny or just an indication of moral uncertainty regarding the other gender in the dataset itself. Overall, the language models that we trained learned the traditional biases associated with users in the Subreddit r/AITA. This indicates that BERT can be highly susceptible to influence from the majority voters regardless of whether that vote is objective or not. This indicates that BERT can learn these biases, some of which are harmful, very easily. Further work can be done in this area by using different language models like XLNet since different language models are trained on different data. An interesting point of research would also be models like GPT-2 and GPT-3 which have been explicitly trained on Reddit data and whether this would bias their text generation.

As an additional future step, we would scrape more data from the Subreddit belonging to the MU class, to increase its representation in our dataset (and improve the class balance). An interesting direction would also be to factor in the temporal component of the Reddit posts and assess if the classifier behaves differently as the years progress - perhaps reflective of the introduction of newer and/or removal of older biases amongst the Redditors themselves? We could also develop a more objective moral classifier, which would be detect ethical acceptability in its truest sense, without ending up replicating human biases. This could be done by employing bias mitigation and dataset de-biasing algorithms in our classification pipeline.

References

[1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and Philip W. Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813v1*, 2002.

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

[3] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30:1731–1741, 2017.

[4] Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. The moral choice machine: Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society. Palo Alto (California): Association for the Advancement of Artificial Intelligence*, 2019.

[5] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[6] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[7] Siqi Liu. Sentiment analysis of yelp reviews: A comparison of techniques and models. *arXiv preprint arXiv:2004.13851*, 2020.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[9] Andrei Mircea. Machine learning to tell you whether you are an asshole or not, 2019.

[10] Elle O'Brien. Aita for making this? a public dataset of reddit posts about moral dilemmas, 2020.

[11] Andreea Salinca. Business reviews classification using sentiment analysis. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 247–250. IEEE, 2015.

[12] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin A. Rothkopf, and Kristian Kersting. BERT has a moral compass: Improvements of ethical and moral values of machines. *CoRR*, abs/1912.05238, 2019.

[13] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

[14] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[15] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13230–13241, 2019.

8. Appendix

8.1. Reproducibility Checklist

- The details regarding the dataset (such as train-validation-test split, batch sizes, etc.) can be found within this report.

- The computing resources used are also mentioned earlier in this report. The runtime for each trained model was roughly 4 hours for each model.

- This is a downloadable link to a folder containing our trained models and results : `https://www.dropbox.com/sh/w3czqgnkfh2d44l/AABIIbi9NRhxyxxrquGJwWgta?dl=0`

- The Readme file in the GitHub repository contains instructions on how to run the necessary code to reproduce the results obtained.