Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 267–276
Marseille, 20-25 June 2022

'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]
[1]King Saud University, Riyadh, Saudi Arabia
[2]University of Manchester, Manchester, United Kingdom
aralhassan@ksu.edu.sa, jinkai.zhang@student.manchester.ac.uk, viktor.schlegel@manchester.ac.uk

Abstract
Natural language processing (NLP) has been shown to perform well in various tasks, such as answering questions, ascertaining natural language inference and anomaly detection. However, there are few NLP-related studies that touch upon the moral context conveyed in text. This paper studies whether state-of-the-art, pre-trained language models are capable of passing moral judgments on posts retrieved from a popular Reddit user board. Reddit is a social discussion website and forum where posts are promoted by users through a voting system. In this work, we construct a dataset that can be used for moral judgement tasks by collecting data from the AITA? (Am I the A*******?) subreddit. To model our task, we harnessed the power of pre-trained language models, including BERT, RoBERTa, RoBERTa-large, ALBERT and Longformer. We then fine-tuned these models and evaluated their ability to predict the correct verdict as judged by users for each post in the datasets. RoBERTa showed relative improvements across the three datasets, exhibiting a rate of 87% accuracy and a Matthews correlation coefficient (MCC) of 0.76, while the use of the Longformer model slightly improved the performance when used with longer sequences, achieving 87% accuracy and 0.77 MCC.

Keywords: moral judgments, AITA subreddit, pre-trained language models.

1. Introduction

In some cases, people might doubt if their behaviour is in line with publicly accepted norms and customs. The increasing number of highly specialised discussion forums on the Internet opened up the opportunity for these people to hear a second opinion about whether they behaved rightly or wrongly in a certain situation. A popular example of these specialised forums is a subreddit on Reddit called AmItheA******?[1] (AITA?), which is dedicated to passing moral judgement on everyday conflicts. A person can open a new thread to describe a situation that led to a conflict between multiple parties and the board community gets the chance to cast one of four different votes with regard to the described situation: not the a****** (NTA), you're the a****** (YTA), no a****** here (NAH) and everyone sucks here (ESH). Voters can also provide an explanation describing how they came to their conclusion, and the majority of votes determines the final verdict.

The discussions in this subreddit can be highly complex; for example, their complexity is reflected in the fact that the stories may involve multiple parties. Furthermore, they usually consist of multiple paragraphs. Thus, passing a rational judgement requires a deep understanding of context. Moreover, this particular type of decision-making requires the ability to link the information conveyed in the post to a specific type of background knowledge, i.e. the set of norms or customs the reader adheres to. This is a core intellectual activity to successfully comprehend a piece of text (McNamara and Magliano, 2009) and is typically not found in other text classification tasks, such as sentiment analysis, which is where a sentiment can often be deduced from having a knowledge of lexical semantics of the expressions used to convey it. The complex nature of these stories has piqued our interest in investigating the potential capability of natural language processing technologies to interpret this challenging cultural and social context. Indeed, it is worth knowing whether an NLP model can predict moral judgment that is carried out by the crowd, which is the case in this subreddit.

In this paper, we aim to utilise the AITA subreddit to investigate whether state-of-the-art NLP models are capable of modelling how moral judgements are passed by the discussion board's community. Specifically, we intend to learn how to predict the most likely moral judgement when given a textual description of the story. To be precise, we want to assess whether we can optimise a model to come to a verdict similar to the one passed by the majority of the commenters. To do this, we will collect a large-scale dataset that includes posts from the corresponding AITA subreddit. We will subsequently use this dataset to train a text classification NLP model and will aim to predict the moral judgement that was passed by the people of the subforum; in other words, we intend to develop the ability to emulate the voting behaviour of discussion-board participants. Note that these endeavours differ in their aim and scope from the recent spark of interest in assessing the moral capacities of language models in a more general sense (Jiang et al., 2021). Even when moral judgements are collected from crowd-workers, their representativeness is questionable at best as moral norms and standards vary widely between different societal contexts. Thus, in this paper, we regard the task strictly as an intriguing problem and caution the reader from drawing conclusions about the general moral capabilities of NLP models based on our findings.

Since we are using Reddit to build our dataset, giving a brief background about the discussion board is necessary. Reddit is a social forum where users can submit posts to one of the topic-specific sub-fora (subreddits) and then other users can write a comment about that post. Readers can upvote a post or a comment if they think it contributes positively[2] to the conversation or is noteworthy in some other way, otherwise they can downvote it. A comment

[1] AITA: https://bit.ly/2QH6CYG

[2] This is the original intention of the voting mechanism, but some people may use different voting criteria that reflect their cultural and personal background.

score can be determined by subtracting the upvotes from the downvotes.

Furthermore, we also aim to assess the ability of the emerging pre-trained language models to pass an ethical judgement in relation to real-life examples of ethical dilemmas. Li et al. (2020) believe that the emergence of the Bidirectional Encoder Representations from Transformers (BERT) is an important turning point in the development of text classification and other natural language processing technologies. Indeed, the BERT-based language model has enhanced the performance of many NLP tasks, including text classification (Devlin et al., 2019). It is based on transformer architecture (Vaswani et al., 2017), which is a simple network structure that is founded on a self-attention mechanism that does not rely on recurrence and convolution. Parallel training can also be performed, which reduces training costs. However, BERT has been trained as a language model to perform two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). This type of model can be utilised by going through a transfer learning process where it can be finetuned in a supervised manner in relation to a specific NLP task. Many improvements have been made to the foundational basis of BERT, such as ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019) and Electra (Clark et al., 2020). In addition, Longformer (Beltagy, Peters and Cohan, 2020) is a new attention mechanism that can effectively deal with the problem of long documents that BERT cannot cope with.

Large-scale, pre-trained models were evaluated using various natural language understanding benchmarks like GLUE (Wang et al., 2019), RACE (Lai et al., 2017) and SQuAD (Rajpurkar et al., 2016). These models performed well and even obtained results that are superior to human performance (Han et al., 2021). Nevertheless, in spite of the great advances of the aforementioned models, there are still some fundamental challenges, especially regarding tasks that require intellectual reasoning. According to Bisk et al., (2020), it is still unknown whether these models can capture deeper concepts of meaning grounded in actions, vision or societal context. It is therefore worth investigating whether these grounded properties can be captured from the surface form of words and their co-occurrence alone, which is the dominating paradigm for obtaining their representations. Another technical challenge mentioned by Han et al. (2021) concerns fine-tuning the pre-trained language models as each downstream task has its own distinct set of parameters that needs to be finetuned in a different way from other tasks. Thus, further experimentation for each task is required to obtain the optimal parameter choice.

The main contributions of this paper include the following elements: proposing the modelling of crowd-sourced moral judgements where we automatically construct a corresponding large-scale dataset by collecting relevant data from a forum, and developing the ability to compare performances and evaluate the success of a number of state-of-the-art, pre-trained language models on this dataset to provide reasonable baselines for the proposed task. This paper is also organised in the following manner. First, we will survey recent work that is linked with our task. Then, we introduce a new AITA corpus and provide a detailed analysis of the presented dataset. After that, we formulate a series of experiments using pre-trained language models, which will be followed by a discussion on the obtained results.

2. Related Work

We briefly surveyed three topics relating to our work to ascertain the latest advances with regard to the area of our study. The first topic is the recent works that have used Reddit as a data source, since we are working on a subreddit. Second, we looked into the NLP-related works on morality, especially those studies that discuss ethical dilemmas that need moral decisions. Finally, we discuss the task of common-sense validation, which has a similar logic to our task as it seeks out prudent and sound judgments that require knowledge that is considered to be common sense. Exploring these subject matters will give us a better understanding of the proposed moral-judgement modelling task.

2.1 NLP Research on Reddit

With the continuous development of online forums, people find it useful and enjoyable to post their experiences and daily dilemmas. Their posts inspired NLP researchers to take advantage of the vast number of posts and use them as a data source. The Reddit platform is one of the online forums that encourages researchers to use their content by providing an official API, which is free and publicly available.

Zellers et al. (2021) mentioned some advantages of using Reddit as a source of data by pointing out that its users are intrinsically motivated and can naturally write complex real texts without external pressure or seeking out a reward. Additionally, active subreddits evolve over time as users keep posting, so new data can be gathered dynamically and optimised models or data analyses can be continuously adapted to new information. Finally, the anonymous decision and voting mechanism provides an opportunity to give an honest and unbiased verdict. According to Ong and Weiss (2000), this anonymity allows posters not to worry about sharing their stories, and commentators do not have to fear retaliation when commenting and passing judgements.

Specifically in relation to the AITA? subreddit, Botzer, Gu and Weninger (2021) finetuned a BERT model to predict whether a user comment passes a positive or negative moral judgement, yet their study did not involve the use of the actual posts and their final verdicts. They also analysed posts on AITA? and other subreddits to investigate the posting behaviour of positive and negative moral users. O'Brien (2020) has experimented with AITA? too by building a dataset and a simple classifier to test the ability of the machine to give a verdict to a post. A synthetic minority oversampling technique (SMOTE) was used on the minority class to balance the dataset, and O'Brien's (2020) logistic regression model can be utilised as a baseline for more advanced experiments and models.

2.2 NLP Research on Morality

Research relating to morality in the field of NLP can be categorised into two tasks: one is called stance detection, which aims to detect whether a given text's attitude towards a particular entity is supportive, oppositional or neutral; it analyses the implied tendency of specific moral topics to

appear in the text (such as the legalisation of abortion) (Mohammad, Sobhani and Kiritchenko, 2017). Another task is related to the concept of moral foundation measurement (Graham, Haidt and Nosek, 2009). The idea behind this concept is to use five sets of moral intuitions to measure people, which are care, fairness, loyalty, authority, and purity. Researchers then use these five moral dimensions to score a specific text, which can subsequently be used to analyse the differences between these dimensions in texts that are used by divergent groups in order to investigate the variances between them (Fulgoni et al., 2016).

Moral judgement differs from the previous two tasks as it aims to analyse whether the behaviour contained in a text meets certain moral standards. It focuses on personal daily life, while the aforementioned two tasks focused on politics or high-level topics, such as abortion, feminism, fairness or cheating. Up to now, there has been some research that has examined moral judgement. For instance, a study by Aletras et al. (2016) was linked to moral judgement but only focused on judicial decision-making. They trained a binary classifier to predict a court's decisions based on text extracted from court cases Their model reached a 79% average level of accuracy. In another paper by Schramowski et al. (2020), they retrained the universal sentence encoder (Cer et al., 2018) to analyse different text sources, such as news, books and material relating to the Constitution and then built a moral choice machine, which aimed to answer sentence level moral choices (e.g. 'should I [action]?') with predicted answers of yes/no.

Meanwhile, Botzer, Gu and Weninger (2021) employed the AITA? subreddit to investigate moral judgments, but the logic behind the building of their dataset differs from ours. They used the comments under the post to predict whether the comment votes for the post would be YTA or NTA, rather than directly predicting the verdict of the post. Another work regarding morality was completed by Hendrycks et al. (2020) who introduced the ETHIC dataset, which includes over 130,000 examples in five different ethical areas (justice, well-being, duties, virtues and common-sense morality). They finetuned four different language models to evaluate the ability of machine learning models to predict moral judgments. Their results show that RoBERTa-large yields the best performance in terms of accuracy. Similarly, Delphi (Jiang et al., 2021) is a learning model that is capable of answering simple ethical questions in the form of three different modes (free-form QA, yes/no QA and relative QA). To achieve that, they constructed the COMMONSENSE NORM BANK dataset, which contains 1.7 million real-life stories with their corresponding moral judgement labels that were gathered via crowd-sourcing. They utilised Unicorn on Rainbow, which is a pre-trained language model by (Lourie *et al.*, 2021) that resulted in a significant improvement over the baseline. Our approach has some similarity with Delphi in terms of using a static ethical dataset to test the success of the model; however, rather than making (rather questionable) general claims about the learnability of moral judgements by neural models, we focus on the capability of predicting majority judgements as they are passed by users of a specific subreddit. Another great efforts by Lourie, Le Bras and Choi, (2021), who created SCRUPLES dataset that has two parts (ANECDOTES and DILEMMAS) where the former was sourced from AITA subreddit, and the latter was a set of pairs, where each pair contains a manually ranked ethical action in order to reflect the real world norms.

2.3 NLP Research on Common-sense Validation

Common-sense validation aims to test whether automated approaches can succeed at tasks that require knowledge that is considered common sense. For example, a sentence that violates common sense is 'John put an elephant into the fridge'. This task is similar to the logic of our proposed task, since it needs to deeply understand the context and to verify whether the conveyed information contradicts some background knowledge, which is ambiguously defined as common sense. Common sense was required in many different tasks, such as machine reading comprehension (Huang et al., 2019), but the task was explicitly formulated as a text classification by task 4 of SemEval-2020 as subtask A (Wang et al., 2021). At this point, participating teams train models to decide which sentences violate the rules of common sense and which ones do not. Participating teams finetuned large-scale pretraining models, such as BERT, RoBERTa and ALBERT, to fit the task. According to Wang et al. (2021), the predictions of the best-ranking models were comparable to human performance, and the most favourable results were obtained by using external common-sense-knowledge resources to help improve the performance. The top ranked team (Zhang et al., 2020), uses K-BERT (W. Liu et al., 2019), which can enhance the performance in a specific domain through the utilisation of knowledge graphs. Then, they modified K-BERT and used a knowledge graph (ConceptNet, Speer, Chin and Havasi, 2017) to help extract further common-sense-based knowledge.

3. Dataset Collection

Motivated by the efforts of O'Brien (2020), we collected a new dataset that contains about 175,000 posts in the period between 1 January 2020 and 15 December 2021. We used their Python code, which is publicly available on GitHub[3]. Some modifications were made to the code to get more specific results, and we used Colab Pro+ notebooks for data collection. However, collecting Reddit posts requires several steps; as a prerequisite, one should have a valid Reddit account. Subsequently, the following steps should be followed.

3.1 Pushshift API

First, we collected the identifications (IDs) of AITA? subreddit posts. We did so by using Pushshift, which is a platform for collecting social media data. It maintains both historical and real-time data and offers flexibility when retrieving large amounts of information; it is not only used for Reddit data, but it can be used for collecting data from other resources (Baumgartner et al., 2020). We employed Pushshift to collect the posts' IDs only, since this would allow us to filter posts by date of publication. Then, we used the official Reddit API to collect the actual content of the posts. Below, Table 1 shows a sample of two collected IDs with their respective Unix timestamp:

[3] https://github.com/elleobrien/AITA_Dataset

[4] https://github.com/praw-dev/praw

ID	Timestamp
qyopl4	1637475174
qyosav	1637475477

Table 1: AITA? ID Samples

3.2 PRAW

After collecting the IDs, we accessed the Reddit API to retrieve the contents of each post and any metadata that was of interest. For that, we used a Python package called Python Reddit API wrapper (PRAW)[4], which allowed us to access the official Reddit API. After that, posts with the corresponding IDs could be retrieved. For each post, we stored a number of fields that are essential to complete our task. Table 2 shows the description of each field; 'Verdict' is one of the most important columns in our dataset, and each row represents the final textual decision that has been selected by the majority of the commenters. This means that each post is annotated with a pre-determined verdict, which eliminates the need for manual annotation. Although the 'Num_comments' column indicates the activity of comments for that post, these comments are not necessarily representative of agreement or disagreement, because they might all be positive or negative. However, the final verdict for this post is determined based on the top voted comment.

Field	Description
ID	The post identifier (e.g. qyopl4).
Timestamp	The Unix timestamp (1637475174).
Title	A short question that starts with AITA? (e.g. AITA for standing up for my father?).
Body	The post's actual text, which is a long paragraph (between two and seven lines).
Edited	Has two values: either 'False', which means not edited after submission or the Unix timestamp that indicates editing time.
Verdict	The decision made about the post, which it can be one of four verdicts: NTA, YTA, NAH or ESH.
Score	Result of subtracting upvotes from downvotes.
Num_comments	Number of comments in each post.

Table 2: Posts' Field Descriptions

One drawback of PRAW is that it cannot retrieve posts between specific dates. We overcame this drawback by using Pushshift as a first step to collect the IDs from the chosen period of time. Another disadvantage of PRAW is that it has restrictive API rate limits in that it takes around one minute to retrieve 30 posts. For that reason, we split the data-collection task between four personal computers (PCs) to accelerate the collection process. Table 3 shows an excerpt from one post and its corresponding verdict.

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. I am the child that my mum relies on for everything, but ignores me the rest of the time. My older sister (34) has always been rude…. etc
Verdict	Not the A******

Table 3: Example of a Post

3.3 Data Cleaning and Pre-processing

In general, social-media text contains a significantly high amount of noise. Other than spelling mistakes and typos, the noise can stem from the sporadic use of punctuation, different letter cases and the omission of stop words. Eliminating such noise will help when trying to achieve a consistent format, saving memory and speeding up the classification process. Nonetheless, for BERT-like language models, it is unclear whether data pre-processing and normalisation yield significant performance gains. Kumar, Makhija and Gupta (2020) suggested that synthetic noise needs to be eliminated to enhance the performance of BERT. In addition, when dealing with Reddit posts, some special pre-processing steps need to be completed to raise the quality of the posts. For that reason, we carried out the following steps to clean and pre-process our dataset:

- Removed posts with lower scores to ensure that posts have received a certain amount of attention;
- Removed posts with a blank body, including [deleted] and [removed] posts;
- Removed [AITA] keyword from the title;
- Transformed all text into lowercase;
- Removed hyperlinks and line breaks (\n character)

Then, by following O'Brien's (2020) method, we merged the verdicts of YTA and ESH into binary class 1, since they both lead to the same positive judgement for that person. The same was done in relation to NTA and NAH by merging them into binary class 0 as they both indicate a negative judgement (see Table 4). Consequently, this turned our task into a binary classification problem.

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

Table 4: Simplified Classes

Finally, we combined the title column with the body column to benefit from the key information provided in the title of the post. After the above data was cleaned and pre-processed, we were left with 110,000 posts.

3.4 Dataset Analysis

In this section, we explore some descriptive statistics and insights from the cleaned dataset. Our final dataset contains eight columns and an additional column that represents the binary class of the verdict. Starting with the body column, we found that the maximum number of words in a post was 1994 word, which is significantly longer than the maximum capacity of BERT-like language models (512 tokens where words can be split in multiple tokens). The average number of words was 368 and more than 75% of posts have a word count of 512 words or fewer, which means that at least 25% of the posts are too long to be processed completely by most of the language models. In terms of frequent words, the most frequent noun in all the posts' bodies is 'Mom', whereas the most prevalent one from posts relating to YTA and ESH verdicts is 'friend', and the most common noun in the posts linked to NTA and NAH verdicts is 'family'.

All Verdicts	YTA+ESH	NTA+NAH
Mom	Friend	Family
Time	Time	Mom
Friend	Mom	Time
Family	Family	Friend
Work	Parents	Parents
Sister	Home	Money
Parents	Wife	Husband
Dad	Money	Job
Home	Husband	Kids
Money	Daughter	Boyfriend
Brother	Boyfriend	School

Table 5: Top-ten Most Frequent Words for Each Verdict

As can be seen from Table 5, these frequent words differ slightly in terms of their rankings from verdict to verdict, but they are all similar and revolve around typical examples of matters in our daily life, as we previously mentioned. Moving on to the verdict column, we calculated the frequency of each verdict as well as the frequency of the binary labels (Table 6 and Figure 1). As we can see, the vast majority of the posts were labelled as NTA.

Verdict	Frequency	Binary Labels	Frequency
YTA	19139	1	24105
ESH	4966		
NAH	7901	0	86613
NTH	78712		

Table 6: Verdict and Binary Label Frequencies

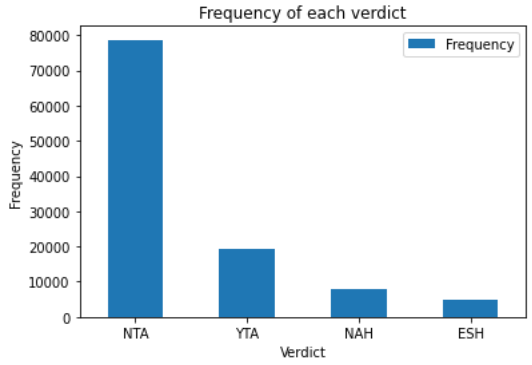

Figure 1: Frequency of each verdict.

3.5 Balancing the Dataset

Figure 1 makes it evident that there is a fair amount of class imbalance as the majority of the posts are labelled as NTA, which might negatively affect the machine learning process. For this reason, we decided to split the resource into datasets with balanced class distributions. Since we do not know exactly which examples are potentially useful samples for the learning process, we randomly under-sampled the majority NTA class, taking this into account by removing very long posts that might exhaust the learning process (i.e. posts with word counts longer than the average number of words). We equalised the number of examples for each class, making each one encompass 24,000, which lead us to a 1:1 class distribution, and we called the resulting balanced subset 'Subset2'. Adhering to the same under-sampling technique, an additional third subset was also created by choosing posts longer than the average word length in order to experiment with the impact of longer sequences. Table 7 describes the three resulting datasets.

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

Table 7: Dataset and Subsets Descriptions

We made all of the dataset and subsets publicly available[5] to give researchers an opportunity to explore other insights into the data.

4. Experiment

State-of-the-art pre-trained language models, based on the Google Transformer architecture (Vaswani *et al.*, 2017) have performed well when finetuned in relation to many downstream tasks. For text classification, models are required to process the full sentence. Encoder models, including BERT, ALBERT, RoBERTA and RoBERTA-large, have a bi-directional attention, which makes them suitable for this task. In this paper, we will adapt each of the four encoder models to our task by finetuning these models with the AITA dataset. In particular, we will investigate the ability of these language models to pass moral judgements on real-life scenarios. Moreover, for longer sequences, we will ascertain the robustness of the Longformer model when handling these sequences.

4.1 Experiment Settings

For all of the classification models, we used the following settings and configurations: codes were implemented using the TensorFlow platform and the models were retrieved from the TensorFlow Hub[6]. We finetuned the models on the NVIDIA Tesla P100 GPU and set the batch size for most of the models to eight or lower to prevent 'out of memory' errors occurring. We then added a classification head for each model. In addition, since we have a binary classification task, we used the binary cross-entropy loss function and the Adam optimiser with a learning rate of 2e-5. To evaluate the performance of the fine-tuned models, we utilised Scikit-learn's train-test-split[7] to randomly split the data, and we reserved 75% of the dataset in each experiment for the training and the remaining 25% for the evaluation.

Table 8 highlights the employed models along with their respective total number of parameters and the corpus size, which each model has been originally pre-trained with.

Model	Total Parameters	Corpus Size
BERT	110 million	16GB
RoBERTa	125 million	160GB
$\text{RoBERTa}_{\text{Large}}$	340 million	
ALBERT	12 million	16GB
Longformer	148 million	77GB

Table 8: Model Specifications

[5] https://bit.ly/AITA_Dataset

[6] https://www.tensorflow.org/hub

[7] https://scikit-learn.org

4.2 Finetuned BERT Model

For the first model, we used 'bert_en_uncased' as a BERT layer, which has been pre-trained in relation to 3.3 billion words. This uncased version is not the case-sensitive iteration of BERT; thus, all of the characters in the posts need to be converted to a lower-case format before tokenisation. This model is restricted to handling up to 512 tokens as the length of the input sequences. For the tokenisation, the WordPiece tokeniser (Wu et al., 2016) was used to break down the words to meaningful subwords, but only if the whole word was not included in the WordPiece vocabulary file, which consists of 30,000 tokens. The resulting tokens went on to be further pre-processed to generate the three essential inputs (i.e. input_word_ids, input_mask and input_type_ids) that are expected by the model. The output of the BERT layer includes two outputs: a pooled output with representations for the entire input sequence, and sequence output with representations for each input token (in context). We only downstreamed the pooled output in our model. Finally, we added a classification head to the BERT Layer and aimed to minimise overfitting on the training set by adding dropout regularisation fixed at 0.4.

In the first experiment, Dataset1 was used with its imbalanced nature, which reflects the real-world distribution, where the majority of posts were judged to be NTA. We also wanted to examine if using more training examples yielded better accuracy or not. Around 25% of Dataset1's posts included sequences longer than a 512-word length and a maximum length of 1994. For that reason, we truncated posts that were longer than 512 tokens and only used the beginning of the post as the input. For the second experiment, the balanced subset was used (Subset2). Here, we did not need to truncate as the maximum token length in this dataset was 512 by design, so we trained the model with full-length sequences. Training in relation to the whole of Dataset1 took six hours (two hours per epoch), whereas the training with regard to subset2 took one hour and 15 minutes per epoch. Both datasets have an average training time of 100 samples per minute.

4.3 Finetuned RoBERTa Model

RoBERTa is an improved version of BERT where the hyperparameters are optimised and the training procedure is improved to upscale the performance. This model has been trained on a diverse array of data, including STORIE, which is tailormade for common-sense reasoning tasks (Trinh and Le, 2018), meaning it is linked to a similar domain to our task. This makes RoBERTa a potentially preferable candidate to be finetuned as it could transfer the learnt knowledge. It uses a variant version of byte pair encoding (BPE) by Sennrich, Haddow and Birch (2016) for text tokenisation, which has a vocabulary size of 50,000. Similar to BERT, RoBERTa accepts up to a 512-word length for each post.

We ran experiments on the RoBERTa-base version for both Dataset1 and Subset2. The sequence truncation method was set to true for Dataset1, and training took around three and half hours for Dataset1 and three hours for Subset2.

4.4 Finetuned RoBERTa-Large Model

In order to investigate the relationship between the size of the model and its performance, we ran two experiments on the large version of RoBERTa. The base version of RoBERTa contains 123 million parameters, whereas the large version holds up to 354 million parameters; consequently, we were not be able to train the data on a standard GPU. For that reason, we used TPUv2, which is available on Google Colab, to accelerate the workload and to handle the large number of parameters.

Two experiments were conducted using this model. Because of the memory limitations, we used our smallest subset (Subset2) for both experiments. For the first experiment, we set the maximum length of the sequences to 512. In the default settings, it takes 18 hours of training for each epoch. To mitigate the slow training process of RoBERTa-large, we employed distributed data parallelism to accelerate the training time to one hour per epoch. In the second experiment, we decreased the batch size to four and kept the maximum length as it is (Max_length=512). This change in batch size reduced the training time in a single GPU to two hours per epoch.

4.5 Finetuned ALBERT Model

ALBERT is the light version of BERT; in particular, it addresses the hardware memory limitation problem by performing two parameter-reduction methods: factorised embedding parameterisation and cross-layer parameter sharing (Lan et al., 2019), which allow us to efficiently optimise even larger language models. The tokenizer used by ALBERT is the SentencePiece tokeniser (Kudo and Richardson, 2018), which was employed to perform subword tokenization with a vocabulary size of 30,000.

We conducted three experiments using the 'albert-base-v2' model. In the first one, we experimented with ALBERT's ability to handle the unbalanced Dataset1; then, another experiment was conducted with regard to its balanced counterpart, Subset2. The training duration for these two experiments was one hour and half an hour, respectively, for a batch size of eight. Meanwhile, the lightweight nature of ALBERT, which is represented by the reduced number of parameters used in this model, allowed us to increase the batch size to 16 with a significant training speed increase to 26 minutes per epoch in a single GPU. This was the fastest training procedure among all of the experiments conducted.

4.6 Finetuned Longformer Model

The previously mentioned pre-trained models have a 512 token length limit, which could result in the loss of significant information for sequences longer than 512 tokens due to sequence truncation (Beltagy, Peters and Cohan, 2020). Longformer, on the other hand, can handle longer sequences without shortening them; indeed, this model can be finetuned to handle up to 4096 tokens. While 75% of posts in our main dataset are shorter than 512 words, it is worth examining the impact of the remaining 25% posts that have sequences that are longer than this to assess the capability of the Longformer model to capture contextual information from long posts when using its modified self-attention mechanism. We used the 'allenai/longformer-base-4096' version for our three experiments. For the first experiment, while utilising a

standard GPU, we set the maximum sequence length for Dataset1 to 1024 and decreased the batch size to four to overcome memory limitations. It took one hour to complete one epoch of training. Likewise, the same experiment was repeated in relation to Subset3, which contains several posts that are longer than 512 words. As a final experiment, we increased the maximum sequence length to 2048 and kept the batch size to four.

5. Experiments' Results

5.1 Evaluation Metric

To establish the performance of the models, we calculated accuracy scores for the binary classification task; accuracy was defined as the ratio of correct classifications to the total number of classifications.

$$\text{Accuracy} = \frac{\text{Correctly classified post}}{\text{Total number of classified posts}}$$

However, accuracy alone is not sufficient to evaluate the classifier's performance because it is highly dependent on the class distribution. So, for the imbalanced dataset, accuracy may produce misleading results that are based on the prediction of the majority class. For that reason, we adopted an additional comprehensive metric that considers the performance of both classes. The MCC metric takes into account all of the confusion matrix values, i.e. true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP}) \cdot (\text{TP}+\text{FN}) \cdot (\text{TN}+\text{FP}) \cdot (\text{TN}+\text{FN})}}$$

This metric has values between -1 and 1 where (-)1 indicates the perfect (anti-)correlation of predictions and expected labels and zero indicates that there is no correlation between the predictions and the ground truth. Thus, to obtain high MCC scores, the classifier must be able to classify both classes correctly (Chicco and Jurman, 2020).

5.2 Results

Table 9 summarises the results of different finetuning experiments that have been carried out in terms of accuracy and MCC. Starting with the unbalanced Dataset1, RoBERTa and BERT performed almost equally well and showed about an 0.1 MCC value, which means that the classifier has drawn little attention to the minority class. In addition, according to the confusion matrix in Table 7, both ALBERT and Longformer failed to recognise any posts from the minority class.

Moving on to subset2, we noticed that ALBERT and RoBERTa performed similarly, yet there was a slight difference that favoured the latter model. Moreover, increasing the batch size from eight to 16 with regard to ALBERT did not improve the results. Interestingly, upgrading to RoBERTa-large did not lead to a significant improvement either. For the last subset, the Longformer improved substantially when used with a balanced subset and slightly outperformed RoBERTa on the same subset.

Dataset	Model	Sequence Length	Batch Size	Training Accuracy	Validation Accuracy	MCC	Confusion Matrix			
							TP	FP	TN	FN
Dataset1	BERT	512	8	0.78	0.78	0.091	203	174	21480	5823
	RoBERTA	512	8	0.78	0.78	**0.098**	131	52	21584	5913
	ALBERT	512	8	0.78	0.78	0	0	0	21584	6096
	Longformer	1024	4	0.78	0.78	0	0	0	21636	6044
Subset2	BERT	512	8	0.83	0.78	0.59	4007	583	5679	2019
	RoBERTA	512	8	0.81	**0.81**	**0.644**	4029	298	5912	2049
	RoBERTA Large	512	8	0.86	0.79	0.6	3882	440	5775	2191
		512	4	0.75	0.77	0.54	4308	1071	5144	1765
	ALBERT	512	8	0.83	0.80	0.623	3704	257	6093	2234
		512	16	0.76	0.79	0.62	3480	92	6258	2458
Subset3	RoBERTA	512	8	0.86	0.87	0.76	4527	63	5964	1492
	Longformer	1024	4	0.87	**0.88**	**0.77**	4698	140	5854	1354
		2048	2	0.87	0.87	0.763	4495	42	5985	1524

Table 9: Finetuned Models' Results

6. Discussion

In relation to the unbalanced dataset, the results show the inability of the models to obtain results that were substantially better than random guessing. They were only capable of predicting the majority class of the dataset; this suggests that data imbalance is a challenge for the introduced task, which future improvements over the introduced baseline will need to take into account.

In relation to Subset2, RoBERTa-large did not outperform the base version in relation to both batch sizes. This is surprising and eludes a simple explanation. A possible reason is that the large model fails to excel over the base model due to the limited size of Subset2. The best performance relating to this grouping of data was achieved by RoBERTa, which exhibited an MCC of 0.64. Meanwhile, regarding Subset3, Longformer was found to have performed slightly better than RoBERTa. This can be attributed to its ability to learn dependencies from the long sequences contained in that particular subset. No significant improvement resulted when increasing the Longformer's maximum sequence length to 2048. However, additional extensive experimentation with a more careful choice of hyperparameters might be required to investigate this further.

Counterintuitively, the same RoBERTa model performed significantly better when trained and evaluated in relation to Subset3, which contained longer posts than Subset2, despite having the same input length, class balance and number of training examples. Simply put, this means that longer posts have more regular patterns than shorter posts and these patterns can be exploited by neural models. Additionally, they must appear at the beginning, because for Subset 3, the input to the model was truncated to the first 512 tokens.

Furthermore, regarding the class imbalance problem, we can infer that using a balanced dataset will yield significantly better results, which raises the need to boost the minority class with more examples. Overall, pre-trained models exhibited good predictions in relation to balanced data, which validates their robustness when capturing the language characteristics of this particular downstream task. However, they are far from mirroring human performance, which by construction is perfect, suggesting that additional research is required to develop better performing models, particularly with regard to unbalanced data.

However, one caveat to our task formulation was that we did not consider commenters' disagreements. The final verdict for each post is assigned based on the highest number of upvotes for a particular verdict, but this decision is not necessarily representative of every voter. One way to alleviate this problem is to treat the task as (structured) regression rather than classification, meaning the models would be optimised to predict the share of votes that went to a label rather than ascertaining the decision voted (only) by the majority. In any case, judging such ethical dilemmas can be difficult for individuals; oftentimes, these judgements are subjective and can be subject to many environmental factors and beliefs. Jiang et al. (2021) acknowledged this by suggesting a thorough moral textbook should be tailored to teaching the machine how to differentiate between right and wrong and pay attention to time, diverse cultures, demographics and beliefs.

7. Conclusion and Future Work

In this paper, we created a publicly available dataset containing posts from the AITA? subreddit in order to evaluate the performance of pre-trained language models, particularly in relation to the task of passing a judgement on real-life dilemmas. Our experiments show that different pre-trained models can succeed at the task to a varying degree, with accuracies ranging between 78% and 88% on average. Given the limitation placed on the long sequence length of 25% of our main dataset's posts, Longformer overcame this limitation and outperformed RoBERTa by a small margin. Overall, our results indicate that making a correct judgement is not a trivial task that can be easily solved by employing out-of-the-box approaches, and it requires larger balanced datasets to extensively supervise the learning of pre-trained models.

Future efforts relating to this subject matter can be summarised by three suggestions. First, in terms of the dataset, we need to boost the minority class performance, by adding many more examples. We can also utilise another balancing technique, such as SMOTE. Second, in terms of classification, this task can be turned to a multilabel classification or regression task, thus taking into account all the verdicts rather than relying on the majority vote. Finally, external ethical knowledge can be used to improve the rationality of verdicts.

8. References

Aletras, N. *et al.* (2016) 'Predicting judicial decisions of the European court of human rights: A natural language processing perspective', *PeerJ Computer Science*, 2016(10), pp. 1–19. doi: 10.7717/peerj-cs.93.

Baumgartner, J. *et al.* (2020) 'The Pushshift Reddit Dataset', (Icwsm).

Beltagy, I., Peters, M. E. and Cohan, A. (2020) 'Longformer: The Long-Document Transformer'. Available at: http://arxiv.org/abs/2004.05150.

Bisk, Y. *et al.* (2020) 'Experience grounds language', *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 8718–8735. doi: 10.18653/v1/2020.emnlp-main.703.

Botzer, N., Gu, S. and Weninger, T. (2021) 'Analysis of Moral Judgement on Reddit'. Available at: http://arxiv.org/abs/2101.07664.

Cer, D. *et al.* (2018) 'Universal sentence encoder for English', *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 169–174. doi: 10.18653/v1/d18-2029.

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*. BMC Genomics, 21(1), pp. 1–13. doi: 10.1186/s12864-019-6413-7.

Clark, K. *et al.* (2020) 'ELECTRA: Pre-training Text

Encoders as Discriminators Rather Than Generators', pp. 1–18. Available at: http://arxiv.org/abs/2003.10555.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, pp. 4171–4186.

Fulgoni, D. *et al.* (2016) 'An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. BT - Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.', pp. 3730–3736. Available at: http://www.lrec-conf.org/proceedings/lrec2016/summaries/1076.html.

Graham, J., Haidt, J. and Nosek, B. A. (2009) 'Liberals and Conservatives Rely on Different Sets of Moral Foundations', *Journal of Personality and Social Psychology*, 96(5), pp. 1029–1046. doi: 10.1037/a0015141.

Han, X. *et al.* (2021) 'Pre-Trained Models: Past, Present and Future', *AI Open*. KeAi Communications Co., Ltd. doi: 10.1016/j.aiopen.2021.08.002.

Hendrycks, D. *et al.* (2020) 'Aligning AI With Shared Human Values', pp. 1–29. Available at: http://arxiv.org/abs/2008.02275.

Huang, L. *et al.* (2019) 'Cosmos {QA}: Machine Reading Comprehension with Contextual Commonsense Reasoning', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2391–2401. doi: 10.18653/v1/D19-1243.

Jiang, L. *et al.* (2021) 'Delphi: Towards Machine Ethics and Norms', (1), pp. 1–42. Available at: http://arxiv.org/abs/2110.07574.

Kudo, T. and Richardson, J. (2018) 'SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing', *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 66–71. doi: 10.18653/v1/d18-2012.

Kumar, A., Makhija, P. and Gupta, A. (2020) 'Noisy Text Data: Achilles' Heel of BERT', pp. 16–21. doi: 10.18653/v1/2020.wnut-1.3.

Lai, G. *et al.* (2017) 'RACE: Large-scale ReAding comprehension dataset from examinations', *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 785–794. doi: 10.18653/v1/d17-1082.

Lan, Z. *et al.* (2019) 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations', pp. 1–17. Available at: http://arxiv.org/abs/1909.11942.

Li, Q. *et al.* (2020) 'A Survey on Text Classification: From Shallow to Deep Learning', 31(11), pp. 1–21. Available at: http://arxiv.org/abs/2008.00364.

Liu, W. *et al.* (2019) 'K-BERT: Enabling Language Representation with Knowledge Graph'.

Liu, Y. *et al.* (2019) 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', (1). Available at: http://arxiv.org/abs/1907.11692.

Lourie, N. *et al.* (2021) 'UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark'. Available at: http://arxiv.org/abs/2103.13009.

Lourie, N., Le Bras, R. and Choi, Y. (2021) 'SCRUPLES: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes', *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15 SE-AAAI Technical Track on Speech and Natural Language Processing II), pp. 13470–13479. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/17589.

McNamara, D. S. and Magliano, J. (2009) *Chapter 9 Toward a Comprehensive Model of Comprehension*. 1st edn, *Psychology of Learning and Motivation - Advances in Research and Theory*. 1st edn. Elsevier Inc. doi: 10.1016/S0079-7421(09)51009-2.

Mohammad, S. M., Sobhani, P. and Kiritchenko, S. (2017) 'Stance and sentiment in Tweets', *ACM Transactions on Internet Technology*, 17(3). doi: 10.1145/3003433.

O'Brien, E. (2020) *AITA for making this? A public dataset of Reddit posts about moral dilemmas*, *DVC*.

Ong, A. D. and Weiss, D. J. (2000) 'The impact of anonymity on responses to sensitive questions 1', *Journal of Applied Social Psychology*. Wiley Online Library, 30(8), pp. 1691–1708.

Rajpurkar, P. *et al.* (2016) 'SQuad: 100,000+ questions for machine comprehension of text', *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, (ii), pp. 2383–2392. doi: 10.18653/v1/d16-1264.

Schramowski, P. *et al.* (2020) 'The Moral Choice Machine', *Frontiers in Artificial Intelligence*, 3(May), pp. 1–15. doi: 10.3389/frai.2020.00036.

Sennrich, R., Haddow, B. and Birch, A. (2016) 'Neural machine translation of rare words with subword units', *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3, pp. 1715–1725. doi: 10.18653/v1/p16-1162.

Speer, R., Chin, J. and Havasi, C. (2017) 'ConceptNet 5.5: An Open Multilingual Graph of General Knowledge', (Singh 2002).

Trinh, T. H. and Le, Q. V. (2018) ‘A Simple Method for Commonsense Reasoning’. Available at: http://arxiv.org/abs/1806.02847.

Vaswani, A. *et al.* (2017) ‘Attention is all you need’, in *Advances in neural information processing systems*, pp. 5998–6008.

Wang, A. *et al.* (2019) ‘Glue: A multi-task benchmark and analysis platform for natural language understanding’, *7th International Conference on Learning Representations, ICLR 2019*, pp. 353–355. doi: 10.18653/v1/w18-5446.

Wang, C. *et al.* (2021) ‘SemEval-2020 Task 4: Commonsense Validation and Explanation’, pp. 307–321. doi: 10.18653/v1/2020.semeval-1.39.

Wu, Y. *et al.* (2016) ‘Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation’, pp. 1–23. Available at: http://arxiv.org/abs/1609.08144.

Zellers, R. *et al.* (2021) ‘TuringAdvice: A Generative and Dynamic Evaluation of Language Use’, pp. 4856–4880. doi: 10.18653/v1/2021.naacl-main.386.

Zhang, Y. *et al.* (2020) ‘CN-HIT-IT.NLP at SemEval-2020 Task 4: Enhanced Language Representation with Multiple Knowledge Triples’, pp. 494–500.