Delft University of Technology, Bachelor Seminar of Computer Science and Engineering

Evaluating catastrophic forgetting of state-of-the-art NLP models for predicting moral values

Florentin-Ionut Arsene[1] , **Pradeep Murukannaiah**[1] , **Enrico Liscio**[1]

[1]TU Delft

Abstract

Personal moral values represent the motivation behind individuals' actions and opinions. Understanding these values is helpful both in predicting individuals' actions, such as violent protests, and building AI that can better collaborate with humans. Predicting moral values is a challenging problem due to the abstract and subjective essence of moral values. With the help of seven Twitter datasets corresponding to different domains, we train state-of-the-art Natural Language Processing models in predicting moral values. An interesting limitation of the models is that they all suffer from catastrophic forgetting. Catastrophic forgetting is the degree to which models worsen their performance on older data after being trained on new data. We conclude that catastrophic forgetting occurs irrespective of the models being trained and can be mitigated by not only training on new data but by training on a combination of old and new data. This is all possible under one assumption: old data is available. We deliver an evaluation of catastrophic forgetting for each model, explain the differences between the models, and suggest possible future work that can be built upon this research.

Keywords: Natural Language Processing, Moral Values, Moral Foundation Theory, Catastrophic Forgetting

1 Introduction

Moral values represent the underlying motivation behind people's opinions, which influence their day-to-day actions. Recent research [14] shows that understanding moral values from opinionated text on social media can be used to predict violent protests. Predicting moral values is also helpful in order to build value-aligned AI [6] that can better operate amongst humans. Due to their subjective essence, it is challenging to estimate personal moral values from text. However, recent Natural Language Processing (NLP) developments facilitate estimating moral values from text.

Recently, Hoover et al. [8] released the MFTC dataset, consisting of "35,108 tweets that have been curated from seven distinct domains of discourse and hand annotated by at least three trained annotators for 10 categories of moral sentiment"[8]. Moral Foundation Theory [17], which theorizes that human morality is primarily dependent upon five different moral foundations, allows estimating moral values from text. The tweets in the dataset are labeled according to the Moral Foundations Theory. This dataset allows for an in-depth analysis of state-of-the-art NLP models in predicting moral values.

A particular characteristic of NLP models is that they suffer from catastrophic forgetting. Catastrophic forgetting, also known as catastrophic interference, is a common weakness of artificial neural networks [5] that occurs when a pre-trained model is trained to perform new tasks. Catastrophic forgetting can be defined as the degree to which the pre-trained model worsens its performance on old tasks after learning to perform new tasks.

In this research, we are going to evaluate the catastrophic forgetting of state-of-the-art NLP models, such as transformers (BERT) [3], optimized text processing frameworks (fastText)[16] and LSTM [15], in predicting moral values. Understanding the catastrophic interference of different state-of-the-art NLP models is helpful when choosing a deep learning model designed to sequentially learn to perform new tasks without forgetting to perform previously learned tasks. For instance, consider a model trained to perform well on the different datasets corresponding to different domains. Given a new dataset corresponding to a new domain, one might ask: how will the model perform on the initial datasets if it is trained on the new dataset?

In order to evaluate catastrophic forgetting of the state-of-the-art models in predicting moral values from text, we sequentially train NLP models on different datasets, mentioned in [9], and evaluate the performance of the models on the initial datasets which were used for pre-training. We deliver a well-documented evaluation of catastrophic forgetting for the NLP models and we show that the degree of catastrophic forgetting can be improved by combining the training set of the old dataset with the training set of the new dataset, allowing the model to learn to perform the new task while still being optimized to perform old tasks.

2 Related Work

Predicting moral values from opinionated text is a challenging problem due to the nature of moral values and the limited available datasets in this area of research.

Hoover et al.[8] released a dataset, consisting of 35 thousand tweets, gathered from Twitter and annotated according to the Moral Foundations Theory. They also present results of various models such as LSTM(Long Short-Term Memory) and SVMs (Support Vector Machines), showing that the dataset allows for the prediction of moral values from text. With the help of this dataset, other researchers show improvements in the predictions of moral values from text, i.e. Leuenen et al. [18] did a cross-domain classification study on moral foundations prediction. None of the research so far analyzes the catastrophic forgetting of the state-of-the-art NLP models in predicting moral values from opinionated text.

Catastrophic forgetting is a significant problem in deep neural networks [5], including Natural Language Processing models. Catastrophic forgetting occurs when a deep neural network that learned to perform a task with optimal parameters tries to perform different tasks. Simply learning a new task can drastically change the optimal weights for previously learned tasks, thus leading the neural network to worsen its performance on old tasks. Many techniques were proposed to avoid catastrophic forgetting, such as combining old training data with new training data, Learning Without Forgetting (LWF) [12], Elastic Weight Consolidation (EWC) [11]. Given old task A and new task B, the main idea is that catastrophic forgetting is overcome by finding the optimal parameter space for both tasks. This can be visualized as the intersection of the two parameters spaces in figure 1.

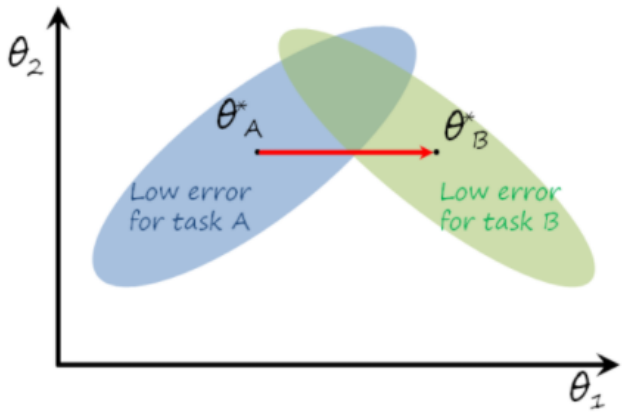

Figure 1: Continual Learning: The highlighted regions represent the optimal parameters spaces for tasks A and B. The intersection of the two regions represents the optimal parameters space that yields low errors on both tasks. Figure taken from [1].

Even though some techniques were proven to mitigate catastrophic forgetting, those techniques were not proven to work in the field of predicting moral values from text. None of the research so far analyzed catastrophic forgetting in models for predicting moral values. We aim to contribute to this line of research by showing a comprehensive evaluation of catastrophic forgetting of state-of-the-art NLP models for predicting moral values.

3 Methodology

In this section we will describe the methodology used to evaluate catastrophic forgetting of the models in predicting moral values. We will first describe the datasets (section 3.1), the pre-processing of the datasets (section 3.2), and the data annotation strategy (section 3.3). Moreover, we will describe the models chosen (section 3.4). In the last section (section 3.5), we will introduce the main topic, catastrophic forgetting, and motivate what experiments are going to be made to answer the research question.

3.1 Datasets

Predicting moral values is a multi-label text classification problem. Given a text input, we aim to label it one or more moral values, according to the Moral Foundation Theory [17]. Table 1 shows the detailed description of the Moral Foundation Theory taxonomy. If the text is detected not to contain moral values, the "non-moral" label is assigned.

Foundation	Definition
Care Harm	This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.
Fairness Cheating	This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy
Loyalty Betrayal	This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
Authority Subversion	This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
Purity Degradation	This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

Table 1: Definitions of moral foundations. Table taken from [4].

We used the Moral Foundations Twitter Corpus [8], recently published and proven to be helpful in training models in predicting moral values. It is a dataset of over 30 thousand tweets corresponding to different socially relevant topics: All Lives Matter, Black Lives Matter, Baltimore Protests, Hurricane Sandy, #MeToo, The 2016 Presidential Elections and hate and offensive speech: the Davidson dataset. Each tweet is annotated by at least 3 different annotators according to the Moral Foundations Theory Taxonomy. Since all the topics with their tweets are sufficiently different, we treat them as different domains. Table 2 shows a description of each of the seven domains.

Corpus	Corpus Description	N
All Lives Matter (ALM)	Tweets related to the All Lives Matter movement	4,424
Black Lives Matter (BLM)	Tweets related to the Black Lives Matter movement	5,257
Baltimore Protests	Tweets posted during the Baltimore protests against the death of Freddie Gray	5,593
2016 U.S. Presidential Election	Tweets posted during the 2016 U.S. Presidential Election	5,358
Hurricane Sandy	Tweets related to Hurricane Sandy, a hurricane that caused record damage in the United States	4,591
#MeToo	Tweets related to the MeToo movement	4,891
Davidson Hate Speech	Tweets collected by Davidson et al. (2017) for hate speech and offensive language research	4,873

Table 2: Description of each of the 7 domains present in MFTC.

Collecting all the tweets from the public API of Twitter is impossible at the moment of writing, as many of the tweets either violate Twitter's policies or have been deleted by the authors. Figure 2 shows the number of available tweets compared to the number of tweets presented in the MFTC dataset [8], only half of them being available. In order to retrieve the full dataset, we contacted Hoover et al. [8], who sent it to us.

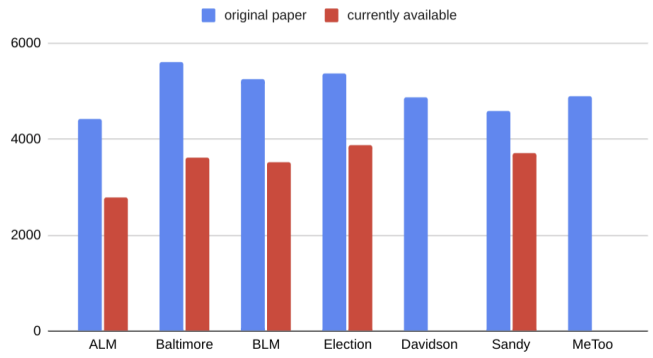

Figure 2: Tweets still available for retrieval using the public API. Figure taken from [7].

3.2 Data pre-processing

The raw data needs to be processed due to its nature. As it is collected from Twitter, it contains many usernames, URLs, and simbols that are not relevant to the context of the data and limit the models from performing to their full potential. Vecerdea et al. [20], who analyzes the performance and transferability of word embeddings, gives an in-depth explanation of the pre-processing strategies. In our case, we are using only the fifth strategy, which was proven to be the best strategy for BERT. Strategy 3 achieves slightly better results than strategy 5 for LSTM; however, we aim for a fair comparison among the models due to the nature of our experiments. A short description of the different strategies can be seen in table 3.

3.3 Data annotation

Every tweet in the dataset had three or more annotations. The annotation strategy that we used was to label each tweet with the values that were present in at least 50% of the annotations.

Strategy	0	1	2	3	4	**5**
Lowercase	✗	✓	✓	✓	✓	✓
No usernames, urls and numbers	✗	✓	✓	✓	✓	✓
No # symbol	✗	✓	✓	✓	✓	✓
No punctuation	✗	✓	✓	✓	✓	✓
No emojis	✗	✓	✓	✗	✓	✗
No stopwords	✗	✗	✓	✓	✓	✓
Lemmatization	✗	✓	✓	✓	✗	✗
Word segmentation	✗	✗	✗	✓	✓	✓
Spell correction	✗	✗	✗	✓	✓	✓
Emoji to words	✗	✗	✗	✗	✗	✓

Table 3: The six pre-processing strategies. Table taken from [20].

With this annotation strategy, each tweet can be labeled with one or more moral values. Whenever the tweet had no labels, it is labeled as non-moral. The frequency of tweets per label can be visualized in table 4.

3.4 Models

We chose to implement three different NLP models that help us better answer the research question. We chose LSTM as a baseline to compare current developments in moral value classification, fastText as a very fast text processing framework, and BERT as the state-of-the-art model.

LSTM

LSTM (Long Short-Term Memory) [15] is a type of RNN that can learn long-term dependencies. The main feature of LSTM is that it solves the Vanishing Gradient and Exploding Gradient problems, current limitations of simple RNNs, also explained by Bengio et al. [2]. Even though we evaluate state-of-the-art NLP models, LSTM is not state-of-the-art; however, it was used in the most recent experiments by Hoover et al. [8] and Leuenen et al. [18] in predicting moral values from text. Therefore, we choose this as a baseline to compare the other models to it.

Fasttext

fastText [16] is a lightweight, text-processing framework that allows for fast training of supervised and unsupervised Natural Language Processing tasks. Joulin et al. [10] show that fastText performs on par with deep learning models on text classification problems; however, its training time is orders

	ALM	Baltimore	BLM	Election	Davidson	Sandy	#MeToo
Subversion	91	257	303	165	7	451	874
Authority	244	17	276	169	20	443	415
Chearing	505	519	876	620	62	459	685
Fairness	515	133	522	560	4	179	391
Harm	735	244	1037	588	138	793	433
Care	456	171	321	398	9	992	206
Betrayal	40	621	169	128	41	146	366
Loyalty	244	373	523	207	41	415	322
Purity	81	40	108	409	5	56	173
Degradation	122	28	186	138	67	91	941
Nonmoral	1744	3848	1583	2502	4509	1313	1618
Total	4424	5593	5257	5358	4873	4591	4891

Table 4: Frequency of Tweets per Foundation Calculated Based on Annotators' Majority Vote. Figure taken from [8]

of magnitudes faster than deep learning models on a standard CPU, which is easier to set up than on the GPU.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a recent model published by Devlin et al. [3]. It is built on the Transformer [19] and can be easily fine-tuned, by adding one output layer, in order to obtain state-of-the-art results in many different NLP tasks, as explained in [3].

3.5 Catastrophic Forgetting

In this research, we investigate the catastrophic forgetting of the selected models on the seven datasets/domains that are included in the MFTC dataset. In order to evaluate the catastrophic forgetting of the models and understand how they will behave in a real-world application (outside the datasets that we are using), we create experiments that mimic real-world scenarios. All scenarios assume we have a pre-trained model on six datasets, and we receive a new dataset on which we want to train the model. We define the degree of catastrophic forgetting as the difference between the model's performance on the pre-training data before receiving the new dataset and after receiving the new dataset.

No Training

This scenario assumes the pre-trained model on six datasets receives the new datasets that it wants to learn. However, we do not train the model on the new data. Thus, this scenario causes no catastrophic forgetting and allows us to use this scenario as a baseline to compare the other scenarios.

Fine Tune

This scenario represents the pre-trained model that receives a new dataset and cannot access the old data. In this scenario, we train the model on the new dataset and evaluate whether it is sufficient to do so. This represents the scenario when the data cannot be stored for many reasons: privacy, storage space, and it gives an insight into how the models would behave in the real world. We expect catastrophic forgetting to occur here.

Train All

This scenario assumes the pre-trained model has access to old data. After the model receives the new data, the model trains on a combination of the new data and a quantity of the old data that has an equal size to the new data. Thus, we expect the degree of catastrophic forgetting to be relatively low, as the model learns to perform both old and new tasks simultaneously.

4 Experimental Setup

Each result table shows the scores of each experiment per domain. The columns represent the domains chosen as the new domain which the model tries to learn, while the remaining six domains represent the old domains, which the model was pre-trained on. The rows consist of the types of experiments that were run. The experiments were run according to the previously explained scenarios: "No Training", "Fine Tune", "Train All". Table 5 shows an example of a results table. The results under "Domain1" show the performance of the model on the six domains that were used for pre-training, namely "Domain2", "Domain3", "Domain4", "Domain5", "Domain6", "Domain7", after receiving "Domain1" as the new domain. The conclusion that can be drawn from those results is that the model suffers significantly from catastrophic forgetting in the "Fine Tune" scenario. At the same time, there is no catastrophic forgetting in the "Train All" scenario. The "No training" scenario represents the model's performance on the initial domains without training on new data.

Experiments	Domain1	Domain2	Domain3	Domain4	Domain5	Domain6	Domain7
No training	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Fine Tune	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Train All	0.9	0.9	0.9	0.9	0.9	0.9	0.9

Table 5: Catastrophic forgetting example.

The experiments are repeated seven times, each domain being, in turn, the new domain that the model is being trained on

and the remaining datasets being the pre-training data, therefore running all the possible combinations of experiments. Each experiment is run with k-fold cross-validation, with k chosen as 10.

4.1 Metrics

The F1-score is a standard measure used for evaluating binary classification problems. It is the harmonic mean between the precision and recall(See equation 4.1). However, there are three types of F1 scores in multi-label classification problems that can be used, namely micro, macro and weighted.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 * \frac{precision * recall}{precision + recall} \quad (1)$$

While the micro F1 score is calculated using the global numbers of True Positives, False Negatives, and False Positives, not considering the imbalance of labels, macro F1 and weighted F1 are an aggregation of the F1 scores per each label. However, the weighted F1 is highly influenced by the majority labels, yielding high results when the F1 scores for the majority labels are high. On the other side, the macro F1 is not highly influenced by the majority classes, all weighing the same. The macro F1 represents the sum of F1 scores for each label. This will penalize the minority labels more than the micro and weighted.

Since there is an imbalance in the classes of labels throughout the domains, we want the models to perform well on all the labels, not just the majority ones. Therefore, the metric that we are using for evaluating the experiments is the macro F1-score.

5 Results

5.1 LSTM

Table 6 shows the results of the experiments ran on the LSTM model. It can be observed that the performance of the model on the pre-training data is represented by a macro F1 score between 0.4 and 0.5 (in the "No training" row). The total training time of an individual experiment is between 10 and 20 minutes.

The Fine Tune experiments show that the performance of the model on the pre-training data drops significantly when the model is sequentially trained on new domains: 26% decrease in macro F1 score when Baltimore is the new domain, 37% decrease when Davidson is the new domain. Overall, the model worsens its performance on pre-training data after being trained on new domains.

The Train All experiments work under the assumption that the model has access to pre-training data. This is helpful, as the model can train on new domains and keep the performance on old domains by combining a small part of the pre-training data with the new data. It can be observed that catastrophic forgetting barely occurs, seeing a decrease in the macro F1 score on the pre-training data of at most 2%.

5.2 fastText

Table 7 shows the results of the experiments that were ran with the model developed with fastText. It can be observed that the performance of the model on the pre-training data is represented by a macro F1 score between 0.53 and 0.60 (in the "No training" row). The total training time of an individual experiment is between 0 and 5 minutes.

Experiments	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
No training	0.46	0.52	0.39	0.49	0.47	0.47	0.46
Fine Tune	0.38	0.26	0.39	0.12	0.40	0.39	0.37
Train All	0.46	0.48	0.43	0.48	0.48	0.45	0.46

Table 6: Catastrophic forgetting for LSTM.

The Fine Tune experiments show that the model's performance on the pre-training data drops significantly when the model is sequentially trained on each dataset; the decrease in the performance ranging from 18% (Baltimore) to 6% (BLM). Overall, the model worsens its performance on pre-training data, but it can be concluded that it suffers less from catastrophic forgetting than LSTM.

In the Train All experiments, it can be noticed that catastrophic forgetting barely occurs, seeing a decrease in the macro F1 score on the pre-training data of at most 2%, similar to LSTM.

Experiments	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
No training	0.56	0.58	0.53	0.57	0.57	0.59	0.57
Fine Tune	0.49	0.40	0.47	0.42	0.51	0.45	0.50
Train All	0.54	0.56	0.54	0.55	0.56	0.55	0.56

Table 7: Catastrophic forgetting for fastText.

5.3 Bert

Table 8 shows the results of the experiments that were ran with BERT. It can be observed that the performance of the model on the pre-training data is represented by a macro F1 score between 0.62 and 0.68 (in the "No training" row).The total training time of an individual experiment is between 90 and 110 minutes.

The Fine Tune experiments show that BERT suffers from Catastrophic Forgetting similarly to fastText. The table shows that the highest decrease in performance on pre-training data occurs when the target dataset is Davidson: 24%.

In the Train All experiments, it can be noticed that we can mitigate catastrophic forgetting. We can see a decrease in the macro F1 score on the pre-training data of at most 2%, similar to LSTM and fastText.

Experiments	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
No training	0.66	0.68	0.62	0.67	0.67	0.68	0.67
Fine Tune	0.57	0.49	0.55	0.43	0.62	0.61	0.61
Train All	0.64	0.67	0.61	0.65	0.65	0.66	0.65

Table 8: Catastrophic forgetting for BERT.

5.4 Transferability

While we only analyzed the performance of each of the models on pre-training data throughout the three experiments, it

is also worth analyzing the performance of the models on the new domains that are used for training. For each model, we averaged the model's performance on the new domains, repeating this for all types of experiments. Table 9 shows the performance of each model on new domains, on average, for each experiment. It is worth noticing that the models have the lowest performance when they are not trained on new domains without any catastrophic forgetting. On the other hand, the models achieve the highest performance when they are simply fine-tuned but with a high degree of catastrophic forgetting. However, we noticed that catastrophic forgetting could be mitigated with the "Train All" experiment. Therefore, we can conclude that the "Train All" experiment is an excellent middle ground that achieves low catastrophic forgetting and allows the model to perform well on the new data. Dondera et al. [4], who analyzes the transferability of the three models on the MFTC datasets, gives a more in-depth evaluation of the transferability that we briefly mentioned in this paper.

Experiments	LSTM	fastText	BERT
No training	0.29	0.34	0.45
Fine Tune	0.40	0.47	0.54
Train All	0.37	0.40	0.53

Table 9: Performance of the three models on the new domain in macro F1 score, averaged over 7 domains for each type of experiment.

5.5 Discussion

Catastrophic forgetting can be observed to occur irrespective of the chosen model; however, the degree of catastrophic forgetting is the worst for LSTM. With the help of experiments, we showed that catastrophic forgetting could be mitigated by combining new data with a quantity of pre-training data of the size of the new data. The model learns to perform both tasks simultaneously. That, however, works under the assumption that there is access to old pre-training data.

BERT shows the highest macro F1 scores; however, it takes the most amount of time to train. Whereas fastText achieves slightly worse results than BERT, the difference in their training times is significant enough to take into account when choosing between the two models. LSTM also takes a reasonable amount of time to train; however, the results are the worst out of the three models.

6 Responsible Research

6.1 Ethical Implications

This work contributes to the research field of predicting moral values from opinionated text. Moral values are involved in all aspects of our lives, and understanding these values from text can be used for both positive and negative intentions. Furthermore, the consequences of research can be diffused over time and space, and once new technology becomes public, it is difficult to stop its consequences. Therefore, it is helpful to be aware of the possible consequences of one's work.

There are many positive aspects of being able to understand moral values from opinionated text. For example, building a model that can understand moral values can further help understand humans, making it easier to build AI to better collaborate with humans. Another example of a positive impact is that it can help authorities prevent dangerous situations, for example predicting violent protests. However, like most technology, predicting personal moral values can also have disadvantages. For instance, understanding the moral values from text can arguably yield private information about one's identity without their consent.

The research process consists of two components that might stir ethical discussions: the code and the dataset. The code was built using open-source libraries such as Pytorch, Transformers, Pandas and is freely available on Github [13]. The dataset consists of 35k tweets and was made available to us by the authors that released the dataset [8]. Since some tweets do not align with Twitter's policies and some were tweets were removed by the authors, they cannot be found online anymore. To ensure that we do not harm anyone's privacy, we remove the authors' names from the data and preprocess it before using it for our experiments.

6.2 Reproducibility

Reproducibility is one of the default methods used by researchers to prove the correctness of the experiments and improve their debugging workflow. However, when building deep neural networks, reproducibility becomes challenging to achieve since a deep neural network might use millions of different parameters. Neural networks can also be developed to run on GPUs or TPUs, which are hard to control fully.

The reproducibility of our results is one of the key factors that make our work reliable. Reproducibility aims to facilitate easy debugging and validate our work. In order to enable reproducibility, we need to make sure there is free access to both the data on which we ran experiments and the code used for the experiments. Since the data used for experiments are not publicly available, the only way to access it is to request it from the authors of [8]. On the other hand, the code used for the experiments of this research is publicly available on Github [13]. In order to obtain similar results to ours, the same seeds need to be used. Since the deep neural network that runs on the GPU causes some randomness, running the experiments with the same seed will not yield the same result but a very close result.

To reproduce our results, it is helpful to understand the environment and libraries used to run our experiments. All the experiments were run on the HPC clusters at TU Delft. We made use of the available GeForce RTX 2080 Ti GPUs to speed up the computations. The main libraries that were used to develop and run the experiments can be found in the Github repository of this research [13].

7 Conclusions and Future Work

Personal moral values represent a crucial factor in understanding people's opinions and motivations. Predicting such moral values from text can further help understand people better and contribute to building beneficial AI that can collaborate well with humans. To build such AI, a thorough

understanding of the models is required. Features such as catastrophic forgetting, transferability, explainability are all of high importance and can help make an educated choice of the available models. We developed three main NLP models during this research that helped us understand catastrophic forgetting: LSTM, fastText, and BERT. Besides analyzing the occurrence of catastrophic forgetting for each model, we also understood other underlying features of the models:

- BERT outperforms all other models in terms of macro F1-score; it is easily configurable, the main drawback being that it takes a long time to train.
- fastText is the fastest out of all, achieves slightly worse results than BERT; the main disadvantage is that it is hard to configure and might require third-party implementation for certain use cases.
- LSTM achieves the lowest macro F1-score out of all, takes reasonable time to train, and is easily configurable for various experiments.

Choosing a model for any Natural Language Processing task should be done under careful consideration. The requirements of the tasks should be identified before making a choice: time to train, results, types of experiments. Thus, showing different features of different NLP models and an evaluation of catastrophic forgetting for each model will further help future research.

We evaluated the catastrophic forgetting of LSTM, fastText, and BERT. We showed that catastrophic forgetting could be improved similarly for each model by mixing a quantity of data from old domains and data from new domains. This allows the models to learn to perform well on new domains and keep their performance on old domains. This improvement, however, only works under the assumption that the data used for pre-training the model is available. We observed that experiments that assume the old data is not available, prove that the models suffer from catastrophic forgetting. A suggestion of open question that can be answered in the future is: Could EWC [11], or LWF [12] be used to overcome catastrophic forgetting of state-of-the-art NLP models in predicting moral values? Answering this is useful in understanding whether we can overcome catastrophic forgetting under the assumption that we do not have access to old data due to privacy, storage space, or other reasons.

8 Acknowledgements

This research has been performed as the final stage for the Bachelor of Computer Science and Engineering at Delft University of Technology.

I want to express my gratitude to our weekly supervisor, Enrico Liscio, and responsible professor, Pradeep Murukannaiah, for the guidance and availability during these virtual-only times. I would also like to thank my teammates and friends, Andrei Geadau, Alin Dondera, Dragos Vecerdea, and Ionut Constantinescu, for all the help and support during this research. Without them, this research would not have been as fun and enjoyable as it was. Moreover, I would also like to thank my family and friends for supporting me throughout this process.

References

[1] http://www.lherranz.org/2018/08/21/rotating-networks-to-prevent-catastrophic-forgetting/, 2018.

[2] Y. Bengio, P. Frasconi, and P. Simard. The problem of learning long-term dependencies in recurrent networks. In *IEEE International Conference on Neural Networks*, pages 1183–1188 vol.3, 1993.

[3] 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.

[4] Alin Dondera. Estimating transferability of state-of-the-art models in predicting moral values. 2021.

[5] R French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, apr 1999.

[6] Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, 2020.

[7] Andrei Geadau. Performance analysis of the state-of-the-art nlp models for predicting moral values. 2021.

[8] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[9] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[11] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[12] Zhizhong Li and Derek Hoiem. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, dec 2018.

[13] Enrico Liscio, Andrei Geadau, Florentin Arsene, Alin Dondera, Dragos Vecerdea, and Ionut Constantinescu. https://github.com/enricoliscio/nlp-for-values-CSE3000, 2021.

[14] Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396, 2018.

[15] C. Olah. Understanding lstm. http://colah.github.io/posts/2015-08-Understanding-LSTMs, 2015.

[16] Armand Joulin Tomas Mikolov Piotr Bojanowski, Edouard Grave. fasttext, 2016.

[17] Ain Simpson. *Moral Foundations Theory*, pages 1–11. Springer International Publishing, Cham, 2017.

[18] A. F. van Luenen. Recognising moral foundations in online extremist discourse: A cross-domain classification study. Master's thesis, Uppsala University, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[20] Dragos Vecerdea. Moral embeddings: A closer look at their performance, generalizability and transferability. 2021.