Constructing a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Version of June 22, 2023

Jeongwoo Park

Constructing a Pluralist Moral Sentence Embedding Space using Contrastive Learning

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Jeongwoo Park
born in Seoul, South Korea

Interactive Intelligence Research Group
Department of Intelligent Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Constructing a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Author: Jeongwoo Park
Student id: 4773543

Abstract

Moral values influence humans in decision-making. Pluralist moral philosophers argue that human morality can be represented by a finite number of moral values, respecting the differences in moral views. Recent advancements in NLP show that language models retain a discernible level of knowledge in deontological ethics and moral norms of society. However, a model which can only decide either right or wrong cannot fully understand the diverse moral perspectives of humans.

We propose a moral sentence embedding space, which can encompass moral differences, through the state-of-the-art Contrastive Learning framework. We evaluate the moral embedding space both intrinsically and extrinsically via three tasks: classification, moral similarity, and visual analysis. We show that our moral embedding space understands the characteristics of each moral value. Our results also highlight that moral rhetoric is seldom explicit in the text, emphasizing the necessity of additional information such as moral labels.

Thesis Committee:

Chair:	Prof. Dr. Catholijn M. Jonker
University supervisor:	Dr. Pradeep K. Murukannaiah, Faculty EEMCS, TU Delft
University supervisor:	Ir. Enrico Liscio, Faculty EEMCS, TU Delft
Committee Member:	Dr. Odette Scharenborg, Faculty EEMCS, TU Delft

Preface

This thesis paper marks the end of my five-year journey as a computer science student. There have certainly been many ups and downs during my Thesis, but I can confidently tell that I am proud of my accomplishments so far.

I would like to express my gratitude to all those who have supported me throughout my thesis. First of all, I am very grateful to my daily co-supervisor, Ir. Enrico Liscio, for giving me excellent academic support, motivation, and also constructive feedback throughout the past 9 months. I also thank my daily supervisor, Dr. Pradeep Murukannaiah, and my thesis advisor, Prof. Dr. Catholijn M. Jonker for providing me with a wonderful opportunity to grow as a researcher. I would like to thank Dr. Odette Scharenborg, for her valuable contribution as a committee member.

Furthermore, I want to thank my friends and family for their unwavering support throughout my journey. Lastly, special thanks to Dirk who always kept me strong and confident.

Jeongwoo Park
Delft, the Netherlands
June 22, 2023

Constructing a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Jeongwoo Park
Delft University of Technology, the Netherlands

Abstract

Moral values influence humans in decision-making. Pluralist moral philosophers argue that human morality can be represented by a finite number of moral values, respecting the differences in moral views. Recent advancements in NLP show that language models retain a discernible level of knowledge in deontological ethics and moral norms of society. However, a model which can only decide either right or wrong cannot fully understand the diverse moral perspectives of humans.

We propose a moral sentence embedding space, which can encompass moral differences, through the state-of-the-art Contrastive Learning framework. We evaluate the moral embedding space both intrinsically and extrinsically via three tasks: classification, moral similarity, and visual analysis. We show that our moral embedding space understands the characteristics of each moral value. Our results also highlight that moral rhetoric is seldom explicit in the text, emphasizing the necessity of additional information such as moral labels.

1 Introduction

Moral values are the glue that binds society together (Lin et al., 2017; Haidt, 2012). It is crucial for fast-growing, autonomous artificial intelligence systems to align with human moral values (Gabriel, 2020). Several frameworks that link AI and ethics have been proposed, but the majority of them does not incorporate individual moral differences or the existence of moral value conflicts (Telkamp and Anderson, 2022). Although there is always a desire for AI systems to behave ethically, consensus on what elements constitute the morally right action is lacking (Awad and Levine, 2020; Telkamp and Anderson, 2022).

Haidt and Joseph (2004) introduced the Moral Foundations Theory (MFT), where they explain morality using an analogy with taste. They draw parallels between five taste receptors—sweet, sour, salty, bitter, and umami—which are used to please tongues across different cuisines, and the five moral taste receptors. MFT states that people have five innate moral foundations on which they base their moral decisions. In other words, we can use MFT to consider the moral and value differences between individuals (Telkamp and Anderson, 2022). MFT is a well-known theory which has been used together with NLP research over time (Araque et al., 2020; Kobbe et al., 2020; Liscio et al., 2022a; Alshomary et al., 2022). In this paper, the Moral Foundation Twitter Corpus (MFTC), an MFT-based dataset, is chosen as the key data source of the experiment. It consists of 35k tweets annotated based on the theory. Each foundation is a vice/virtue pair.

Prior research has often relied on classifiers to recognize moral rhetoric from textual discourse (Lin et al., 2017; Alshomary et al., 2022; Hendrycks et al., 2021; Liscio et al., 2022a). Although the classifiers have shown their effectiveness, they are still focused on a single NLP task, classification. To realm more NLP tasks, another approach should be taken, which is Sentence Embedding. Sentence embedding is a numerical representation which encapsulates knowledge from textual data. Sentence embedding models are often fine-tuned to perform diverse Natural Language Processing (NLP) tasks such as text classification, response selection, information retrieval, semantic text similarity, and intent discovery (Reimers and Gurevych, 2019; Henderson et al., 2020; Chen et al., 2022). Considering the high practicality of sentence embedding, there exists a significant potential for exploring its applications in capturing morality within textual data. Hence, this paper employs a sentence embedding framework to comprehend moral values present in textual discourse.

Recently, there has been an increasing interest in adopting Contrastive Learning (CL) to improve sentence embeddings (Gao et al., 2021; Wu et al., 2022; Zhang et al., 2022). The CL objective entails

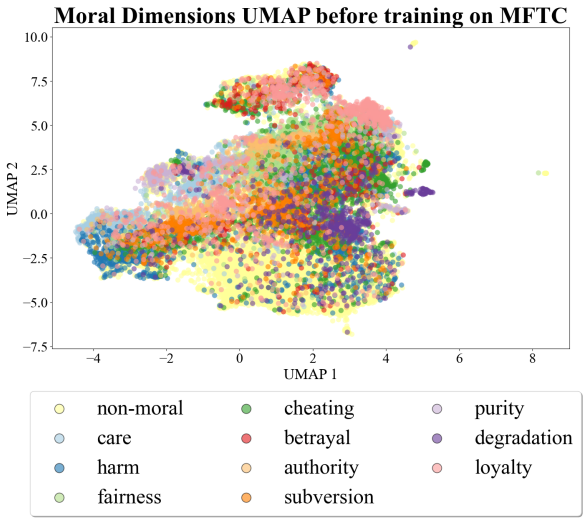

Figure 1: UMAP plot of Moral Foundations before training on the MFTC train set. Plotted using the MFTC train set.

regularization of the embedding space by pulling positive (i.e., semantically similar) sentences closer while distancing negatives (i.e. semantically dissimilar) (Zhang et al., 2022). Gao et al. (2021) introduced *Simple Contrastive Learning of Sentence Embeddings* (SimCSE) which shows remarkable effectiveness in capturing semantic property. In this work, we adopt SimCSE and enrich it with MFT. We also observe the influence of label information on generating a CL-based moral sentence embedding space.

Schramowski et al. (2022) identified that existing pre-trained language models (e.g. BERT) preserve knowledge about deontological choices and even moral norms of society. Nevertheless, as depicted in Figure 1, the dimension-reduced vectors do not form any clusters, indicating that the vectors generated by SimCSE prior to training on MFTC hardly have any insight into MFT. Considering the UMAP plot and the baseline classification result in Section 5.1, it is difficult to claim that large language models understand the characteristics of moral foundations.

In this study, we propose the novel mapping of text to the moral embedding space via supervised and unsupervised CL. This mapping can capture and reflect the moral difference between individuals. We evaluate the moral embedding space with extrinsic evaluation by testing the performance of the embedding space in a common downstream task, classification. We also assess and interpret the moral relationship between each vice and virtue moral value with an intrinsic evaluation which consists of a visual methodology, namely a UMAP plot and Moral Similarity Task (MST).

Our contribution is twofold: (1) we present a method to map the MFT taxonomy to an embedding space using CL, and (2) we show that additional information, beyond the text itself, is crucial to generate an effective moral embedding space.

Our study is significant because an MFT-based embedding space can recognize the characteristics of moral values. By exploring and exploiting the embedding space, we believe that it can make valuable contributions to various NLP tasks.

2 Background

Sentence Embedding with Contrastive Learning Over the past few years, social media platforms generated a large amount of data including text, images, videos and graphs (Wang et al., 2020). To extract knowledge from the textual data, sentence embedding can be used to obtain the vector representation of a sentence. Many state-of-the-art pre-trained language models (PLM) are proposed including BERT, SBERT, and ELMo (Devlin et al., 2019; Reimers and Gurevych, 2019; Peters et al., 2018).

Contrastive Representation Learning aims to learn an embedding space by pushing positive sentence pairs (i.e. semantically similar pairs) closer while pulling apart negative sentence pairs (i.e. semantically dissimilar pairs). SimCSE (Gao et al., 2021) is a text-based CL framework which supports both *supervised* and *unsupervised* approaches. *Supervised* SimCSE applies the CL objective on the positive and negative labelled instances. *Unsupervised* SimCSE creates the positive pair by applying the dropout twice on the query sentence, and it takes another sentence from the same mini-batch as the negative. SimCSE aims to provide more uniformly distributed and better alignments of positive pairs in the embedding space by integrating the CL objective.

Moral Foundations Theory and Corpus MFT is chosen as a basis of our moral embedding space. The MFT is a well-established theory in the field of social science and psychology. The MFT includes five foundations, and each foundation is bipolar, with each pole representing either a virtue or vice as shown in Table 1.

Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020), a corpus annotated based

Foundation	Definition
Care/ Harm	Prescriptive concerns of caring for others/ Prohibitive concerns of not harming others
Fairness/ Cheating	Prescriptive concerns about fairness and equality/ Prohibitive concerns of not cheating or exploiting others
Loyalty/ Betrayal	Prescriptive concerns of prioritizing one's inner-circle/ Prohibitive concerns of not betraying one's inner-circle
Authority/ Subversion	Prescriptive concerns of respecting authority and tradition/ Prohibitive concerns of not subverting to authority or tradition
Purity/ Degradation	Prescriptive concerns of keeping the purity of sacred entities/ Prohibitive concerns on contaminating those

Table 1: The five foundations of MFT.

on MFT, is chosen to generate the moral embedding space. MFTC is a collection of 35,108 tweets which consists of 7 domains: All Lives Matter (ALM), Baltimore Protest (BLT), Black Lives Matter (BLM), hate speech and offensive language (DAV) (Davidson et al., 2017), 2016 presidential election (ELE), MeToo movement (MT), and hurricane Sandy (SND). The tweets are annotated by multiple annotators based on 11 vice-virtue labels and *non-moral*. The final annotation was decided by a majority vote policy, and if there is no majority label, *non-moral* is assigned.

3 Generating the Embeddings

To create a robust sentence embedding space, We fine-tune the SimCSE model on the MFTC via both supervised and unsupervised settings.

Supervised SimCSE requires a fixed format of a dataset as an input for CL. The dataset should have either a triple of {query instance, positive instance, hard negative instance} or a pair consisting of only two positive instances. We decided to construct the triple format due to its higher performance demonstrated in (Gao et al., 2021). However, in our case, choosing negative instances is not trivial because there is no absolute answer as to what is similar/dissimilar in terms of morality. Thus, we came up with two policies, *within* and *outside*, to simplify this process using MFT labels as shown in Figure 2.

Figure 2 shows the steps to create the supervised dataset using an example. The MFTC train set is divided into half to apply different hard negative policies. Both halves find a positive instance which

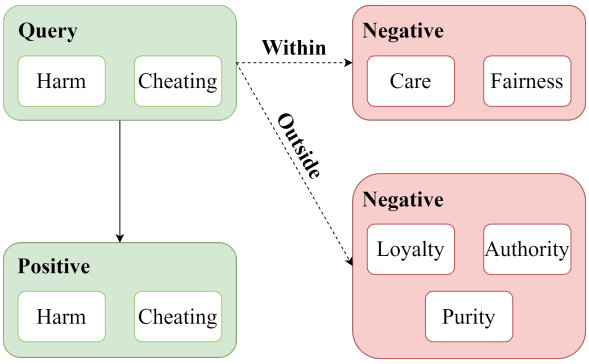

Figure 2: Example of a Supervised Triple formation.

shares the same label. Then each half takes a different policy for choosing the hard negative instance. One finds the hard negative **within** the foundation, and the other finds the hard negative **outside** the foundation. Both **within** and **outside** create the subsets of all the negative labels respectively. Negative labels refer to moral labels excluding the label of query and positive instances. If available, we prioritize the items with more negative labels when choosing the hard negative. For instance, with **outside** policy, we prioritize the tweet with the label "Care" and "Fairness" compared to "Care" only. The same logic applies for **within**. When all the hard negative candidates run out in both scenarios, non-moral items are used as the hard negative. In the last stage, the non-moral items are also used as positive instances while using any foundation to be a hard negative. The final supervised dataset consists of 5304 triples.

The unsupervised dataset does not include any label information. Thus, only tweets are taken to form the unsupervised dataset. SimCSE creates the CL input by applying the methods described in Section 2.

4 Evaluating the Embeddings

Evaluating the high-dimensional embedding poses challenges due to its low interpretability (Senel et al., 2018; Anelli et al., 2022). In this section, we explain how we process the dataset for training and testing. We also discuss how we evaluate our embedding space in two approaches, extrinsic and intrinsic.

The hyperparameter details of SimCSE models can be found in Appendix A.5.

Value	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation	Non-Moral
Train	2176	3269	1870	3068	1736	1736	1294	1816	698	1246	14428
Test	240	359	204	335	183	121	137	196	72	132	1611

Table 2: Label Distribution of MFTC in Train and Test set.

4.1 Dataset Pre-processing

MFTC is pre-processed with the suggested guidelines by (Hoover et al., 2020; Liscio et al., 2022a), and also further described in Appendix A.1.

The train set consists of 90% of MFTC and the test set consists of 10%. The train set is used for generating the moral embedding space, and the test set is left for evaluation. Each dataset obtained 90% (10%) of data from each domain in MFTC to form the train (test) set. We have allocated a high percentage of the dataset to the train set as our research aim is to create a robust moral embedding space. The label distribution can be found in Table 2.

4.2 Extrinsic Evaluation

Extrinsic evaluation is conducted using multi-class multi-label classification.

The two main SimCSE models, *Supervised* and *Unsupervised*, are compared for the classification task. Eger et al. (2019) emphasized the necessity of a simple classifier on top of embeddings to evaluate the embedding space themselves. We decided to use a linear layer (i.e., a fully connected layer), with 1024 input features and 11 output features as a classification head on top of the SimCSE embeddings. 5-fold cross-validation is applied in the classification stage of the SimCSE models.

We choose four models as the baseline models. Two of them are SimCSE models which are not trained on the MFTC train set, *Supervised* SimCSE test-only and *Unsupervised* SimCSE test-only. These models are taken from the baselines provided by (Gao et al., 2021). These two only go through the 5-fold cross-validation, which means only the classifier is trained with 4 folds of the MFTC test set.

The two variants of BERTForSequenceClassification (BFSC) are taken as the baseline models. The baselines are used to simply check the practicality and potential of SimCSE embeddings. BFSC has a classification head on top of the pooled output from BERT (Devlin et al., 2019).

The first variant and the second variant differ in their initial weights. The weights of the first variant are initialized to the BFSC model which is trained on the train set (90% of MFTC). The weights of the second variant are initialized to the regular `bert-large-uncased`, without any knowledge. This second variant is just like the *Supervised* SimCSE test-only and *Unsupervised* SimCSE test-only. Both variants are tested using the 5-fold cross-validation. For simplicity, we refer to the first variant as 'BFSC-all', and the second variant as 'BFSC-test-only'.

For all models, We report the averaged result with Macro and Micro F1-Score as the dataset is imbalanced.

4.3 Intrinsic Evaluation

The intrinsic evaluation focuses on the quality of the embedding space through visual analysis and the moral similarity task using two Moral Foundation Dictionaries.

Moral Similarity Task In NLP, the semantic similarity task is widely used to assess the quality of the language model (Gladkova and Drozd, 2016; Bakarov, 2018). Aligning with our research goal, we adapt the semantic similarity task to Moral Similarity Task (MST). We use two dictionaries, Moral Foundation Dictionary (MFD) 2.0 (Frimer, 2019) and MoralStrength (Araque et al., 2020).

MFD2.0 is an extension of the original MFD with the help of original MFD authors, and MoralStrength is expanded by WordNet synset and evaluated using crowdsourcing. For MoralStrength, each vocabulary is crowd-sourced for the value called *Moral Valence*, the strength with which a vocabulary is expressing the specific moral value. Moral valence ranges from 1 to 9, where a word ranked in the middle signifies neutrality with respect to the specific moral dimension. These are excluded from this task. The words with a strength smaller than 5 are classified as vice of that specific foundation and words with a strength greater than 5 are classified as virtue of that specific foundation. By

including vocabularies with a moderate *Moral Valence*, MoralStrength can possess relatively neutral characteristics when compared to MFD2.0. For example, words with a score 1 and 4 definitely have a strength difference, but are given the same label.

Both dictionaries are employed to compute the moral similarity between each moral value. To accomplish this, the average similarity score between all pairs of words pertaining to the moral values of interest is taken.

UMAP Visual Analysis SimCSE embedding is 1024-dimensional. Hence, relatively little is known about the structure of the embedding space. Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) is a nonlinear dimensionality reduction technique that can be used for visualization (McInnes et al., 2020). McInnes et al. (2020) claims that UMAP preserves both local and most of the global structure in the data with superior run time performance when compared to other dimensionality reduction techniques such as t-SNE. We first get the train set embedding using the two models, the *Supervised* SimCSE baseline and the *Supervised* SimCSE optimal model, and use UMAP to reduce the dimension to 2. We plot them visually to observe the formation of clusters which reflect the effect of training.

5 Results and Discussion

We evaluate the model using the aforementioned three tasks. First, we report the extrinsic evaluation by analyzing the classification results. Then, we perform the intrinsic evaluation via the Moral Similarity Task (MST) and the UMAP plot. MST shows an approximate degree of understanding of MFT by the model. The UMAP Plot visually gives an insight into the status of the embedding space.

5.1 Extrinsic Evaluation: Classification

We compare 6 models, described in Section 4.2. The mean and standard deviation of the results are shown in Table 3. For both Macro and Micro F_1-score, we perform the Wilcoxon rank sum test between the best result and the other results to evaluate whether the two results are significantly different. We highlight in bold the best result and the results that are not significantly different ($p > 0.05$) from the best.

According to Table 3, *Supervised* SimCSE model outperforms *Unsupervised* SimCSE model in terms of the classification task. The rank sum test confirms that the supervised result is significantly different from the unsupervised result. This significant difference indicates that the label information in addition to the moral text is beneficial for our model to learn MFT. The result of BFSC-all is significantly better than both SimCSE models in terms of the Macro-F1 score, which takes into account the label imbalance issue. Consequently, BFSC-all displays proficiency in learning under conditions where labels are imbalanced. Overall, the difference between *Supervised* SimCSE and BFSC-all is trivial. We underscore that the focus of this task is on the validation of the SimCSE embeddings, rather than identifying the most optimal classification model.

Table 3: Classification result for 6 models: Supervised SimCSE and Unsupervised SimCSE, BFSC-all, Supervised SimCSE test-only, Unsupervised SimCSE test-only, and BFSC-test-only. The best-performing models are highlighted. * indicates that no train set is used.

Model	**Micro** F_1	**Macro** F_1
Sup. SimCSE	**68.4** $\pm$ 3.1	56.7 $\pm$ 2.6
Unsup. SimCSE	58.0 $\pm$ 2.9	36.2 $\pm$ 3.4
BFSC-all	**71.0** $\pm$ 1.5	**62.2** $\pm$ 1.1
Sup. SimCSE test-only*	59.4 $\pm$ 3.1	39.4 $\pm$ 3.9
Unsup. SimCSE test-only*	58.4 $\pm$ 3.1	37.1 $\pm$ 3.5
BFSC-test-only*	66.2 $\pm$ 2.4	55.8 $\pm$ 1.2

All test-only models are not exposed to the MFTC train set. BFSC-test-only achieves the highest among these models. We believe that end-to-end learning of BFSC facilitates it to learn how to classify. The result of the *Unsupervised* SimCSE does not improve after training on the MFTC train set. This implies that the Unsupervised SimCSE is not effective at learning the characteristics of MFT. In the meantime, both *Supervised* SimCSE and BFSC-all perform better after training on the MFTC train set, by 15% and 7% respectively. Notably, *Supervised* SimCSE shows a substantial increase in performance, highlighting the promising potential of learning with the CL objective.

5.2 Intrinsic Evaluation: Moral Similarity Task

Figure 3 show the pairwise cosine similarity score between words from each moral value. Figure 3a and Figure 3c show the moral similarity table of *Supervised* SimCSE for MFD2.0 and MoralStrength respectively. Figure 3b and Figure 3d display the result of Unsupervised SimCSE using the same

(a) Moral Similarity Table of MFD2.0 Supervised

	care	fairness	loyalty	authority	purity	harm	cheating	betrayal	subversion	degradation
care	0.58	0.31	0.37	0.32	0.39	0.3	0.19	0.23	0.19	0.22
fairness	0.31	0.48	0.33	0.37	0.33	0.25	0.3	0.28	0.27	0.22
loyalty	0.37	0.33	0.51	0.36	0.38	0.26	0.25	0.33	0.32	0.27
authority	0.32	0.37	0.36	0.48	0.4	0.26	0.26	0.31	0.36	0.27
purity	0.39	0.33	0.38	0.4	0.57	0.27	0.21	0.27	0.31	0.35
harm	0.3	0.25	0.26	0.26	0.27	0.56	0.36	0.36	0.33	0.42
cheating	0.19	0.3	0.25	0.26	0.21	0.36	0.52	0.46	0.41	0.39
betrayal	0.23	0.28	0.33	0.31	0.27	0.36	0.46	0.55	0.51	0.39
subversion	0.19	0.27	0.32	0.36	0.31	0.33	0.41	0.51	0.56	0.41
degradation	0.22	0.22	0.27	0.27	0.35	0.42	0.39	0.39	0.41	0.54

(b) Moral Similarity Table of MFD2.0 Unsupervised

	care	fairness	loyalty	authority	purity	harm	cheating	betrayal	subversion	degradation
care	0.36	0.27	0.3	0.28	0.29	0.3	0.27	0.28	0.27	0.28
fairness	0.27	0.32	0.28	0.28	0.27	0.26	0.28	0.28	0.28	0.27
loyalty	0.3	0.28	0.38	0.31	0.3	0.29	0.3	0.34	0.32	0.29
authority	0.28	0.28	0.31	0.34	0.3	0.28	0.29	0.3	0.31	0.28
purity	0.29	0.27	0.3	0.3	0.33	0.28	0.27	0.29	0.29	0.29
harm	0.3	0.26	0.29	0.28	0.28	0.35	0.29	0.31	0.3	0.3
cheating	0.27	0.28	0.3	0.29	0.27	0.29	0.34	0.34	0.32	0.3
betrayal	0.28	0.28	0.34	0.3	0.29	0.31	0.34	0.42	0.36	0.31
subversion	0.27	0.28	0.32	0.31	0.29	0.3	0.32	0.36	0.39	0.31
degradation	0.28	0.27	0.29	0.28	0.29	0.3	0.3	0.31	0.31	0.33

(c) Moral Similarity Table of MoralStrength Supervised

	care	fairness	loyalty	authority	purity	harm	cheating	betrayal	subversion	degradation
care	0.45	0.3	0.37	0.35	0.35	0.29	0.26	0.28	0.3	0.24
fairness	0.3	0.41	0.33	0.37	0.35	0.3	0.39	0.34	0.35	0.32
loyalty	0.37	0.33	0.45	0.39	0.37	0.3	0.32	0.33	0.35	0.3
authority	0.35	0.37	0.39	0.42	0.39	0.31	0.34	0.33	0.36	0.32
purity	0.35	0.35	0.37	0.39	0.46	0.33	0.34	0.34	0.36	0.4
harm	0.29	0.3	0.3	0.31	0.33	0.47	0.35	0.4	0.38	0.41
cheating	0.26	0.39	0.32	0.34	0.34	0.35	0.43	0.4	0.38	0.39
betrayal	0.28	0.34	0.33	0.33	0.34	0.4	0.4	0.43	0.42	0.42
subversion	0.3	0.35	0.35	0.36	0.36	0.38	0.38	0.42	0.44	0.4
degradation	0.24	0.32	0.3	0.32	0.4	0.41	0.39	0.42	0.4	0.52

(d) Moral Similarity Table of MoralStrength Unsupervised

	care	fairness	loyalty	authority	purity	harm	cheating	betrayal	subversion	degradation
care	0.33	0.28	0.3	0.29	0.29	0.29	0.28	0.29	0.3	0.28
fairness	0.28	0.31	0.28	0.28	0.28	0.28	0.3	0.29	0.29	0.28
loyalty	0.3	0.28	0.32	0.29	0.29	0.28	0.28	0.3	0.3	0.28
authority	0.29	0.28	0.29	0.31	0.29	0.29	0.29	0.3	0.3	0.28
purity	0.29	0.28	0.29	0.29	0.31	0.29	0.28	0.29	0.3	0.3
harm	0.29	0.28	0.28	0.29	0.29	0.33	0.29	0.31	0.32	0.3
cheating	0.28	0.3	0.28	0.29	0.28	0.29	0.32	0.3	0.31	0.29
betrayal	0.29	0.29	0.3	0.3	0.29	0.31	0.3	0.32	0.32	0.3
subversion	0.3	0.29	0.3	0.3	0.3	0.32	0.31	0.32	0.34	0.3
degradation	0.28	0.28	0.28	0.28	0.3	0.3	0.29	0.3	0.3	0.31

Figure 3: Moral Similarity Tables on MFD2.0 and MoralStrength.

data.

According to Figure 3a and Figure 3b, *Supervised* SimCSE displays a more prominent color along the diagonal compared to the *Unsupervised* SimCSE, meaning that the similarity between words belonging to the same moral value is the highest. The same can be found in Figure 3c and Figure 3d as well. This indicates that the embedding space can recognize the morality conveyed in the text. Figure 3a shows a more significant contrast between vice and virtue words compared to Figure 3c and the diagonal line is also more intense. We can infer that the embedding space is more capable of discerning the moral properties of MFD2.0 compared to MoralStrength due to the characteristics of the dictionaries mentioned in Section 4.3. Defining A-B as the relationship between A and B, we also observe that the similarity scores of vice-vice (top-left) and virtue-virtue (bottom-right) are greater than the similarity of virtue-vice values (top-right and bottom-left). This phenomenon is expected as the language models are aware of what is right and wrong (Schramowski et al., 2022). Furthermore, the similarity of vice-vice is higher than the similarity of virtue-virtue for both MFD2.0 and MoralStrength. We can interpret that the characteristics of virtue words are more distinguishable than the characteristics of vice words.

Figure 3b and Figure 3d show the results from the unsupervised model. Overall, both figures do not have a significant diagonal, which means that the unsupervised models do not understand the characteristics of moral foundations, contrary to the supervised models. Again, the diagonal of MFD2.0 is more noticeable compared to the one of MoralStrength with the same reasoning.

Lastly, the positive result of the supervised moral similarity tables also manifests a potential to utilize

the supervised moral embedding space across various MFT-based corpora. While the classification task uses only MFTC, MST shows a successful application of other MFT corpora, the dictionaries. We believe this is the green light to use the moral embedding space outside the trained corpus.

Appendix B.5 contains further statistical analysis regarding MoralStrength.

5.3 Intrinsic Evaluation: UMAP Plot

Figure 4 shows the dimension-reduced plot of the Supervised moral embedding space using the MFTC train set. We observe a solid cluster for each vice and virtue moral value. This is a significant improvement compared to Figure 1.

Virtue values are located on the bottom left of the plot, while vice values are located on the top right of the plot. The embedding space clearly shows the separation between virtue and vice. Moreover, the moral values within the foundation are often forming a symmetry. We can assume that the embedding space understands the bipolar characteristics of MFT for some moral foundations, indicating the relationship between those two moral values.

Non-morals are spread throughout all the clusters, but also a lot of them are scattered between vice and virtue clusters. Although a 2-dimensional plot cannot reflect the whole embedding space, the considerable difference before (Figure 1) and after the training implicates the positive result of the CL-based embedding space.

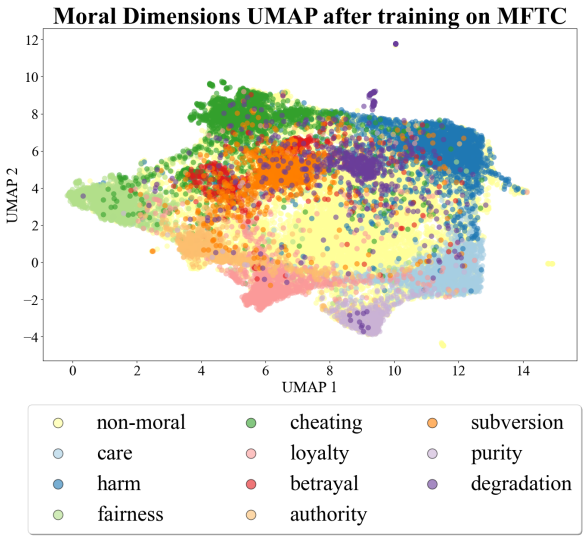

Figure 4: UMAP plot of Moral Foundations after training on the MFTC train set. Plotted using the MFTC train set.

6 Related Work

In this section, we thoroughly examine previous research. We explore different methods to extract values from textual data, multiple moral datasets, and also other contrastive learning sentence embedding frameworks.

6.1 Identifying Moral Values from Text

Previous work on estimating MFT values from the text was done in either supervised or unsupervised methods.

The most common approach in the unsupervised method is to utilize value lexicons. The first version of the Moral Foundations Dictionary (MFD) was presented by Graham et al. (2009), which is the essential backbone of many lexicon-based value identifications. MFD is comprised of lemmas for vice and virtue for each foundation. Multiple extensions were suggested to improve the value lexicons in MFD (Frimer, 2019; Rezapour et al., 2019; Araque et al., 2020; Kobbe et al., 2020; Hopp et al., 2020).

Nevertheless, using only word-level lexicons to estimate values certainly have limitations such as ambiguity of natural language and restricted range of lemmas (Hulpuș et al., 2020). Hulpuș et al. (2020) suggests a new direction of projecting the MFD lexicon on knowledge graphs (KGs), bringing several benefits: reducing ambiguity and identifying moral entities and concepts. Asprino et al. (2022) proposed frame-based value reasoner, a tool based on a frame semantics approach together with various KGs, including an improved ValueNet. Priniski et al. (2021) uses FrameAxis (Kwak et al., 2021) to obtain a moral embedding space. However, the embedding space is constrained to one domain, and as the author mentioned, evaluation has several limitations. Thus, our approach aims to overcome the limitations by(1) using the corpus including more than one topic, and (2) exploring new evaluation methodologies.

Supervised methods are mostly based on classification (Johnson and Goldwasser, 2018; Hoover et al., 2020; Alshomary et al., 2022). In order to train and evaluate the classifier, an annotated dataset according to a value taxonomy is required. Alshomary et al. (2022), Kobbe et al. (2020) and Liscio et al. (2022a) have used a BERT-based classifier on MFT-based datasets.

To the best of our knowledge, this is the first work on constructing a MFT-based embedding

space using a supervised approach. By leveraging the labeled dataset, MFTC, and a Contrastive Learning sentence embedding space, SimCSE, we provide a moral embedding space capable of capturing the diversity in moral perspectives among individuals.

6.2 Datasets with Morality

Besides MFTC, there is another corpus based on MFT, Moral Foundations Reddit Corpus (MFRC). MFRC adopted a revised version of theory (Trager et al., 2022). Here, Fairness is split into two foundations, Equality and Proportionality, resulting in 6 foundations in total. MFRC consists of 16,123 Reddit comments drawn from 12 different subreddits, and these 12 subreddits can be classified into three buckets: US politics, French politics, and Everyday moral life.

There are other morality datasets based on different value taxonomy that can be used for NLP applications. Kiesel et al. (2022) proposed a dataset of 5270 arguments across 4 countries using Schwartz theory (Schwartz et al., 2012). Qiu et al. (2022) developed a human value dataset with social scenarios organized by Schwartz values. Hendrycks et al. (2021) introduced the dataset with contextualized scenarios about justice, deontology, virtue ethics, utilitarianism, and commonsense moral intuitions.

Among these options, we decided to utilize MFT due to the many precedent work developed upon it and the large dataset annotated with moral foundations.

6.3 Contrastive Sentence Embedding

In addition to SimCSE, there are many other state-of-the-art contrastive learning-based models. DeCLUTR, inspired by Deep Metric Learning, is a universal sentence embedding technique that does not require labeled training data (Giorgi et al., 2021). DeCLUTR also proposes a sampling technique to produce positive instances and negative instances. Moreover, Chuang et al. (2022) proposed DiffCSE which is an equivariant contrastive learning model. DiffCSE is insensitive to certain types of augmentations and sensitive to "harmful" types of augmentations.

7 Conclusion

Certain social events may not universally fall into binary categories of ethics, right and wrong. We want the AI agent to understand and accept diverse moral perspectives beyond the binary definition of morality. Hence, we propose a method to map text to the embedding space based on the theory acknowledging the value pluralism. We conduct a thorough assessment of a moral embedding space via both extrinsic and intrinsic evaluation. The extrinsic evaluation demonstrates that moral label information enhances the performance of the embedding space. We emphasize that this moral embedding space is capable of many tasks while providing comparable performance to a state-of-the-art classifier. The intrinsic evaluation indicates that the moral embedding space recognizes the characteristics of moral values via the similarity task and visual analysis.

Our approach improves the moral understanding of the language model from the basic binary notion of right and wrong to a value pluralistic theory, MFT. We also provide a new guideline to generate a moral embedding space with a state-of-the-art contrastive learning language model. Moreover, our moral embedding space can be substantial to many applications. The embedding space with a proper head can support value-aligned agents in interactive narratives as an MFT-based value prior (Ammanabrolu et al., 2022). It can also play an important role in recognizing moral rhetoric from diverse social issues related to abortion, terrorism, and politics (Sagi and Dehghani, 2014). The embedding space can also improve identifying context-specific morality (Liscio et al., 2022b).

For further research, we recommend expanding this moral embedding space to other MFT corpora or other value theory datasets, and also reflecting on multiple annotators' opinions when building the embedding space i.e. using all crowd annotations directly or using both ground truth labels and information about disagreement (Uma et al., 2022).

8 Ethical Considerations And Limitations

With the remarkable performance of large language models (LLM), it gets appealing to explore LLMs for their moral reasoning (Jin et al., 2022). The complexities associated with responding to morally-charged situations pose a challenge not only for humans but also for LLMs (Jin et al., 2022). Hovy and Spruit (2016) addresses *dual-use* problem where a system developed for a certain purpose leads to unintended negative consequences. For instance, since liberals and conservatives pursue different moral foundations (Graham et al., 2009),

the moral embedding space can be misused to identify and discriminate against people with certain political standpoints. Although any researcher is neither encouraging *dual-use* nor fully responsible for this, it is a good practice to take *dual-use* into consideration to prevent any potential side-effect if possible.

There are also cases where the ethics-related NLP models may extend beyond moral judgments and take part in non-moral statements, including political stances, religious prescriptions and medical advice (Talat et al., 2022). Furthermore, the application of our embedding space to particular domains, such as the legal field, requires cautious deliberation (Leins et al., 2020).

We discuss several ethical considerations and limitations regarding MFTC. First of all, MFTC is composed of English tweets, subjects centered in the United States, leading to demographic bias (Hovy and Spruit, 2016). This can hardly represent every moral standard across different cultures. However, we believe that our suggested methods and evaluation schemes should still be applicable to culturally diverse datasets as well. Moreover, the moral embedding space generated from this study is not unbiased. MFTC consists of tweets related to 7 topics, and the dataset can potentially include more tweets from people with certain perspectives. As Liang et al. (2020) mentioned, *post-hoc debiasing* can be applied to the current moral embedding space, which also prevents any retraining of the sentence embeddings. Lastly, MFTC shows a low annotator agreement (Hoover et al., 2020). This means choosing the majority-voted label as the gold label may enforce the domination of the majority opinions, suppressing the minority. We believe that applying a *strong perspectivist* approach, keeping all the annotations into the subsequent training of a model, can improve the practicality of the embedding space (Basile et al., 2021). Furthermore, the integration of diverse opinions has been known for its potential in the classification tasks (Sudre et al., 2019; Akhtar et al., 2020; Campagner et al., 2021).

We address three primary concerns for the evaluation methodology. First of all, the classification task employs one specific classifier. While other classifiers may yield different results, we emphasize that the purpose of this task is to validate the quality of the embedding space, rather than achieving the highest score. Secondly, moral foundation dictionaries created by the non-WEIRD society, could reveal new strengths and weaknesses of the embedding space. Further investigation can be undertaken to assess whether the embedding space captures the properties of morality when tested on culturally diverse dictionaries. Lastly, the purpose of UMAP is to assist in the intrinsic evaluation of the embedding space, and it cannot be used as a sole evaluation tool. The main benefit of employing UMAP is to easily observe the status of the embedding space and the impact of the training. Thus, additional inspection is required for a detailed geometric analysis of the embedding space.

Last but not least, while our supervised CL dataset generation approach strives for effectiveness and efficiency, we believe that further investigation into CL dataset generation holds potential for improvement not only for identifying moral values but also for other domain applications.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Vito Walter Anelli, Giovanni Maria Biancofiore, Alessandro De Bellis, Tommaso Di Noia, and Eugenio Di Sciascio. 2022. Interpretability of bert latent space through knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information; Knowledge Management*, CIKM '22, page 3806–3810, New York, NY, USA. Association for Computing Machinery.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184.

Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovi. 2022. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 33–41, Dublin, Ireland and Online. Association for Computational Linguistics.

Edmond Awad and Sydney Levine. 2020. Why we should crowdsource ai ethics (and how to do so responsibly).

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *ArXiv*, abs/1801.09536.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing.

Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790.

Minhua Chen, Badrinath Jayakumar, Michael Johnston, S. Eman Mahmoodi, and Daniel Pressel. 2022. Intent discovery for enterprise virtual assistants: Applications of utterance embedding and clustering to intent mining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 197–208, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the evaluation of sentence embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, Florence, Italy. Association for Computational Linguistics.

Jeremy A Frimer. 2019. Moral foundations dictionary 2.0.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, and Brian Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96:1029–46.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133:55–66.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari,

Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11:1057–1071.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2020. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53:232–246.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Ioana Hulpuș, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. 2020. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online). Association for Computational Linguistics.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems*, volume 35, pages 28458–28473. Curran Associates, Inc.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Ying Lin, Joe Hoover, Morteza Dehghani, Marlon Mooijman, and Heng Ji. 2017. Acquiring background knowledge to improve moral value prediction. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah. 2022a. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022b. What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1).

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

J. Hunter Priniski, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu, and P. Jeffrey Brantingham. 2021. Mapping moral valence of tweets following the killing of george floyd.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA. Association for Computational Linguistics.

Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Soc. Sci. Comput. Rev.*, 32(2):132–144.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268.

Shalom Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem dirilen gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103:663–88.

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(10):1769–1779.

Carole H. Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D. Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, Rolf H. Jäger, and M. Jorge Cardoso. 2019. Let's agree to disagree: Learning highly debatable multirater labelling. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 - 22nd International Conference, Proceedings*, volume 11767 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 665–673. Springer. 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019 ; Conference date: 13-10-2019 Through 17-10-2019.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Jake Telkamp and Marc Anderson. 2022. The implications of diverse human moral foundations for assessing the ethicality of artificial intelligence. *Journal of Business Ethics*, 178.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Lili Wang, Chongyang Gao, Jason Wei, Weicheng Ma, Ruibo Liu, and Soroush Vosoughi. 2020. An empirical survey of unsupervised text representation methods on Twitter data. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 209–214, Online. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. InfoCSE: Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States. Association for Computational Linguistics.

A Experimental Details

To make sure our research is reproducible, we share the experiment settings and hyperparameter settings.

A.1 Data Processing

We preprocess the tweets by removing URLs, emails, usernames and mentions. Next, we employ the Ekphrasis package [1] to correct common spelling mistakes and unpack contractions. Finally, emojis are transformed into their respective words using the Python Emoji package [2]. Additionally, there are some independent tweets with duplicated content. Some of them had different labels, which is not desired. They also bring difficulty in generating the supervised CL dataset. Thus, repeated instances of distinct tweet annotations are reduced to one instance by applying another majority vote.

.

A.2 Hyperparameters

To select the most optimal combination of hyperparameters of SimCSE, we perform a grid search. Table A1 and Table A2 show the hyperparameters that were compared, highlighting in bold the best-performing option based on the F1-Scores of the classification result. We used these hyperparameters for every experiment in this paper for consistency. If a parameter is not present in the table, the default value supplied by the framework [3] was used.

Table A1: Hyperparameter settings for training Supervised SimCSE

Hyperparameters	**Options**
Model name	sup-simcse-bert-large-uncased
Max Sequence Length	**64**, 128
Epochs	**2**, 3, 5
Batch Size	16, **32**
Learning Rate	5×10^{-5}
Temperature	0.01, 0.05, **0.1**
Pooler	**cls**

The time taken for *Supervised* SimCSE hyperparameter search is roughly 6-7 hours, and the time taken for *Unsupervised* SimCSE hyperparameter search is approximately 15-16 hours.

Hyperparameters for the classifier are mentioned in Table A3. Classifiers did not go through special hyperparameter tuning as the aim of the research is not focused on the quality of the classifier. Default values and common values chosen in practice were used for the classifier.

Table A2: Hyperparameter settings for training Unsupervised SimCSE

Hyperparameters	**Options**
Model name	unsup-simcse-bert-large-uncased
Max Sequence Length	**64**, 128
Epochs	**1**, 2, 3
Batch Size	16, **32**
Learning Rate	3×10^{-5}
Temperature	0.01, **0.05**, 0.1
Pooler	**cls**

Table A3: Hyperparameter settings for Linear Classifier

Hyperparameters	**Options**
Max Sequence Length	64
Epochs	10
Batch Size	16
Learning Rate	0.01
Dropout	0.1
Loss function	Binary Cross Entropy

We base the hyperparameters of our baseline models according to the hyperparameter combinations suggested by Liscio et al. (2022a), which uses the same corpus and the same model. The number of epochs was set to 10, as the linear classifier uses 10 epochs. The learning rate was optimized as the experiment environment changed.

Table A4: Hyperparameter settings for BERTForSequenceClassification

Hyperparameters	**Options**
Model name	bert-large-uncased
Max Sequence Length	64
Epochs	10
Batch Size	16
Optimizer	AdamW
Learning Rate	**2e-5**, 5e-5
Loss function	Binary Cross Entropy

A.3 Computing Infrastructure

- PyTorch: 1.13.0
- Huggingface's Transformers: 4.2.1
- SimCSE: 0.4
- NVIDIA A40 GPU
- CUDA 11.6

[1] https://github.com/cbaziotis/ekphrasis
[2] https://pypi.org/project/emoji/
[3] https://github.com/princeton-nlp/SimCSE

A.4 Random Seeds

In our experiments, we ensure that the same train test splits are used across different runs of each experiment. Further, to control for any randomness throughout code execution, we fixed the random seeds in the following libraries to 42 where it is necessary:

- Python random.seed
- NumPy (numpy.random.seed)
- PyTorch (torch.manual_seed)
- Tensorflow (tensorflow.random.set_seed)

A.5 Artifacts Used

We primarily use two different types of artifacts, data and model.

MFTC is a collection of 35,108 tweets annotated based on MFT (Hoover et al., 2020). MFTC can be downloaded [4], and used under Creative Commons Attribution 4.0 license. MFD2.0 (Frimer, 2019) can be freely accessed[5]. Lastly, MoralStrength Version 1.1 (Araque et al., 2020) can be found online [6] and also used under GNU Lesser General Public License v3.0.

SimCSE is a sentence embedding model which uses contrastive learning (Gao et al., 2021). SimCSE can be used under MIT license[7]. BERT (Devlin et al., 2019) is selected as a baseline model to compare with SimCSE. The license of BERT is Apache License 2.0 [8].

B Extended Results

In this section, we extend the result shown in the main report.

B.1 Training Time Comparison for Optimal models

Table B1 presents the time performance of classification. We took the most optimal embedding model to measure the time performance. As test-only models do not train with the MFTC train set, the first values are all 0. *Supervised* SimCSE takes significantly less total time for the training process than *Unsupervised* SimCSE and BFSC. *Unsupervised* SimCSE is likely to take more time as it takes in all the tweets and creates a positive instance by itself, while *Supervised* SimCSE gets a formatted input which discards some tweets as described in Section 3. Considering the small difference in the final F_1-score, there is clearly a performance advantage in using SimCSE embeddings. Besides the fast performance, *Supervised* SimCSE embedding space can be used in more diverse scenarios compared to BFSC which is limited to a classification function. Supervised embedding space is fixed, with any top layer, the embedding space can manage more tasks.

Table B1: Time Performance Comparison between three models: BFSC, *Supervised* SimCSE and *Unsupervised* SimCSE. First value of SimCSE is embedding training time and Second value is for training and testing classifier.

Model	**Training Time (s)**
Sup. SimCSE	249 + 10
Unsup. SimCSE	493 + 11
BFSC-all	3521 + 327
Sup. SimCSE test-only	0 + 10
Unsup. SimCSE test-only	0 + 10
BFSC-test-only	0 + 313

B.2 Misclassification Error Analysis

We also inspect both (1) the confusion between moral texts and non-moral texts and (2) the confusion between and within the foundations. We look into the following four types of misclassification errors (which add up to 100%) which are addressed in (Liscio et al., 2022a).

Error I A tweet labeled with one or more moral values is classified as non-moral or no prediction.
Error II A tweet labeled as non-moral is classified with zero or more moral values.
Error III A tweet labeled with a moral value is classified with values from other foundations.
Error IV A tweet labeled as a vice/virtue is classified as the opposite virtue/vice within that foundation.

Table B2 illustrates the misclassification error output. Both *Supervised* SimCSE and *Unsupervised* SimCSE mostly encounter Error 1 and Error 2. The two most frequent errors come from distinguishing between moral and non-moral texts which may be caused by the significant class im-

[4] https://osf.io/k5n7y
[5] https://osf.io/xakyw
[6] https://github.com/oaraque/moral-foundations/tree/master/moralstrength/annotations/v1.1
[7] https://github.com/princeton-nlp/SimCSE/blob/main/LICENSE
[8] https://github.com/google-research/bert/blob/master/LICENSE

Table B2: Misclassification Error for all 6 models

Model	Err. 1	Err. 2	Err. 3	Err. 4
Sup. SimCSE	50.5	30.6	17.3	1.60
Unsup. SimCSE	62.9	24.6	11.3	1.15
BFSC all	28.5	36.9	30.7	3.86
Sup. SimCSE test-only*	62.2	24.8	11.6	1.40
Unsup. SimCSE test-only*	65.1	23.2	10.5	1.25
BFSC test-only*	29.3	38.0	29.8	2.89

Table B3: Examples of misclassified texts with their top 2 similar text from *Supervised* SimCSE model. (cosine similarity, label)

	Misclassified Text	Top 1
Err. 1	i d rather be a clueless kid then an ignorant racist coward hiding behind a computer black lives matter (True: cheating, harm, Predicted: non-moral)	it is sad when my own race treats me like i am a stray dog because i will not engage in racism radical black ideology with them all lives matter (0.857, True: non-moral)
Err. 2	all lives matter no single person deserves the cruel and violent acts that are exposed in this world (True: non-moral, Predicted: harm)	too many senseless murders so many innocent people suffering and losing lives wrong on so many levels all lives matter(0.888, True: harm)
Err. 3	thank you for your work i wish had the same compassion and empathy and strength (True: fairness, subversion, Predicted: Care)	this strength compassion leaves me in complete awe i forgive you charleston shooting black lives matter i am same (0.968, True: care)
Err. 4	hey we don ut obey putin or you (True: subversion, Predicted: authority)	trump is our leader we will obey (0.914, True: authority)

balance between moral and non-moral texts. This can also be expected based on Figure B2, which will be discussed later. In the figure, many non-moral instances overlap with moral instances. We believe that adding additional moral instances can improve the performance of SimCSE embeddings. Additionally, *Unsupervised* SimCSE is more confused between morality and non-morality, and *Supervised* SimCSE sometimes gets more confused between foundations. Similar patterns can be found in *Supervised* SimCSE test-only and *Unsupervised* SimCSE test-only models. Like Classification F1-scores, *Unsupervised* SimCSE does not exhibit a substantial difference after training on the MFTC train set. *Supervised* SimCSE makes fewer mistakes in identifying morality from the moral texts, while making more mistakes in recognizing non-morality from the non-moral texts.

The BFSC models show an approximately equal proportion across Error 1, Error 2, and Error 3. Compared to SimCSE, the models are better at distinguishing morality and non-morality, but worse at finding out the correct foundation. Looking at Error 4, the baseline models make more mistakes between virtue and vice within a foundation compared to SimCSE models.

We observe examples of misclassified text for each error, and the most similar sentence to that text in Table B3 to analyze specific cases. Looking at the example of Error 1, it shows that the embedding space recognizes the query text as non-moral as the most similar sentences are non-moral labeled text. In the case of Error 2, the query sentence is annotated by three individuals in three different labels, fairness, care, and harm, before getting assigned as 'non-moral' due to conflicting opinions. Error 2 exemplifies the different perspectives from which the text can be viewed, explaining why the classifier misclassified it. Similarly, Error 3 instance was once labeled as 'care' by one of the annotators. It is not unreasonable to predict the text as care. The misclassified text of Error 4 discusses obedience in a rather strong manner. The other similar examples also show the authority-related context, showing the opposite moral value within the foundation.

B.3 Detailed SimCSE Classification Result

Table B4 and Table B5 show the mean and standard deviation of F1-Score for each moral value. Overall, a common pattern can be found for all models. Cheating and Harm are the easiest vice values to classify, and Fairness and Care are always the easiest virtues value to classify. On the other hand, the Purity foundation is always difficult to identify in all the models. This could be attributed to the fact that there are fewer examples with the 'Purity' label in the dataset. We speculate that increasing the number of instances in other foundations will improve the classification of other foundations as well.

B.4 Foundation Only Embedding

Before conducting experiments with 11 labels, we initially experimented with 6 labels, which are 5 foundations and non-moral, to figure out the possibility of mapping the text to moral foundations.

Table B4: Detailed Classification Result for the best performing SimCSE models (Mean F1, Standard Deviation)

Model	Sup. SimCSE	Unsup. SimCSE
Care	67.9 (5.2)	56.7 (3.7)
Harm	57.5 (4.8)	48.1 (6.7)
Fairness	71.4 (6.3)	50.3 (8.8)
Cheating	66.0 (3.6)	40.1 (7.7)
Loyalty	61.1 (6.0)	36.7 (15.0)
Betrayal	51.0 (9.4)	16.8 (3.3)
Authority	54.9 (10.4)	30.2 (14.1)
Subversion	37.1 (13.1)	16.3 (3.9)
Purity	46.3 (21.8)	14.3 (10.1)
Degradation	32.2 (12.4)	14.6 (13.6)
Non-moral	78.0 (3.7)	73.9 (3.1)

Table B5: Detailed Classification Result for the BFSC models (Mean F1, Standard Deviation)

Model	BFSC All	BFSC Test-Only
Care	70.5 (4.1)	67.0 (3.3)
Harm	64.7 (4.5)	57.9 (4.3)
Fairness	70.8 (7.8)	68.7 (6.1)
Cheating	71.2 (4.5)	64.8 (4.9)
Loyalty	65.4 (4.5)	59.9 (5.2)
Betrayal	55.5 (13.2)	48.2 (9.7)
Authority	59.6 (7.8)	51.5 (12.9)
Subversion	44.8 (10.2)	39.1 (13.5)
Purity	50.1 (8.1)	41.7 (10.7)
Degradation	52.5 (14.0)	38.4 (14.5)
Non-moral	80.3 (2.3)	77.2 (3.5)

Data processing is the same as the 11-label dataset, but the supervised dataset construction is slightly different as vice and virtue from the same foundation are assigned a label of the foundation they belong to. This means if two instances have the same foundation, they can be a positive pair. Furthermore, only hard negatives outside the actual foundation were considered.

Table B6: Hyperparameter settings for Foundation Only *Supervised* SimCSE

Hyperparameters	Options
Model name	sup-simcse-bert-large-uncased
Max Sequence Length	64
Epochs	3
Batch Size	16
Learning Rate	5×10^{-5}
Temperature	0.05
Pooler	cls

The result can be found in Table B8. The default hyperparameters from SimCSE framework are chosen except for the batch size and they are not optimized. The hyperparameters are listed in Table B6 and Table B7. The result is similar to the 11-label model used in the main experiment. As distinguishing between vice and virtue in addition to the moral foundations is more practical, we decided to proceed with the 11-label model.

Table B7: Hyperparameter settings for Foundation Only *Unsupervised* SimCSE

Hyperparameters	Options
Model name	unsup-simcse-bert-large-uncased
Max Sequence Length	64
Epochs	1
Batch Size	16
Learning Rate	3×10^{-5}
Temperature	0.05
Pooler	cls

Table B8: Foundation Only Classification result. 11 Labels refer to the model we discussed in the main paper.

Model	Micro F_1	Macro F_1
Sup. SimCSE Foundation	68.0	56.7
Unsup. SimCSE Foundation	57.5	39.4
Sup. SimCSE 11 Labels	68.4	56.7
Unsup. SimCSE 11 Labels	58.0	36.2

B.5 MoralStrength Statistical Analysis

We investigate the correlation between MoralStrength values and moral embedding similarity via a statistical analysis. Table B9 illustrates the result of the statistical analysis of Figure 3c.

We made an assumption that there is a negative correlation between the two measures, the distance between two MoralStrength scores and the moral similarity between two words. In other words, as the two moral values become more alike (the difference between the two MoralStrength scores becomes small), the moral similarity between them also increases. To explore the correlation between the difference in moral strength values of two words and their moral similarity, we performed a Spearman correlation test. The null hypothesis is that the correlation coefficient is 0. Table B9 shows that every case is a negative correlation as it is expected, but the evidence is weak to reject the null hypothesis (i.e., the p-value is always larger than 0.05).

Although the Supervised moral embedding space has some knowledge of morality, we speculate that the embedding space does not know the specific degree/strength of a certain moral value as it never explicitly learned those details.

Table B9: Spearman Correlation Result of Moral Strength using Supervised Moral Embedding

Value	Avg. Coeff	Avg. p-value
Care	-0.053	0.283
Harm	-0.091	0.182
Fairness	-0.107	0.121
Cheating	-0.070	0.332
Loyalty	-0.100	0.181
Betrayal	-0.125	0.160
Authority	-0.075	0.232
Subversion	-0.132	0.132
Purity	-0.118	0.123
Degradation	-0.052	0.305

B.6 Plotting with UMAP

Figure B2 shows three clusters after training MFTC, vice, virtue and non-moral. Vice means the aggregation of all the vice moral values, and the same applies for Virtue. Compared to Figure B1, it shows a clearer separation between vice and virtue values. Vice and Virtue clusters are less mixed together, and a bigger gap can be found between them. Both figures are plotted using the MFTC train set.

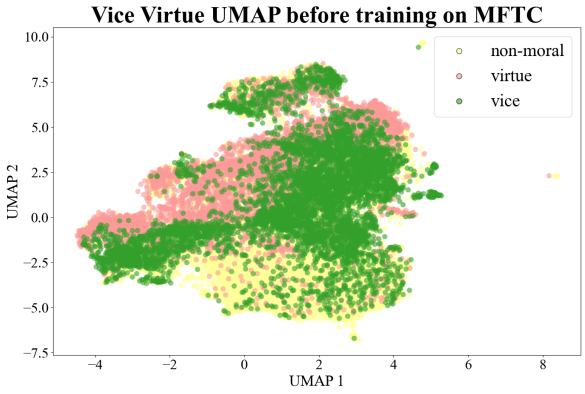

Figure B1: UMAP plot showing only vice and virtue before training on MFTC.

In addition to the train set visualization, we also plot using the MFTC test set. These figures contain more sparse scatter plots than the train set visualization because the test set has fewer data points as described in Section 4.1. The main difference is that the embedding model is not exposed to the test set during the training phase. Hence, it is possible to observe how the clusters are formed for the unseen data.

Figure B3 and Figure B4 show similar patterns as in Figure B1 and Figure B2. The model trained on MFTC can separate the vice and the virtue instances more than the untrained model. Moreover,

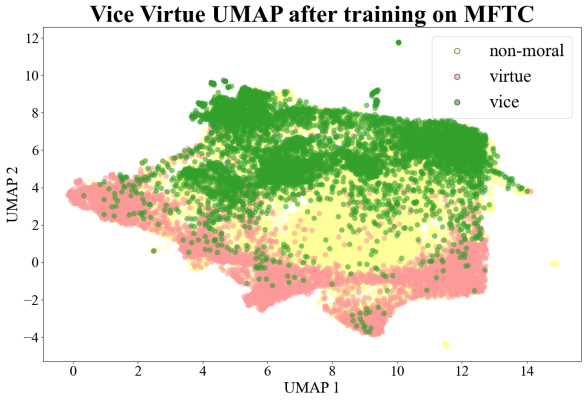

Figure B2: UMAP plot showing only vice and virtue after training on MFTC.

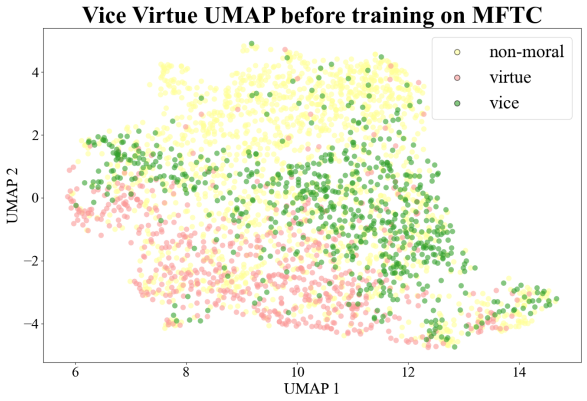

Figure B3: UMAP plot showing only vice and virtue before training on the MFTC train set. Plotted using the MFTC test set.

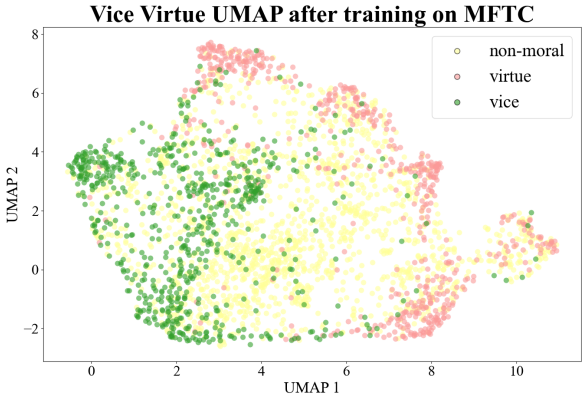

Figure B4: UMAP plot showing only vice and virtue after training on the MFTC train set. Plotted using the MFTC test set.

Figure B6 forms clusters more clearly compared to Figure B5, which is similar to the description in Section 5.3. Overall, the findings from the train set visualization and the test set visualization share many commonalities, confirming the moral knowledge of the Supervised SimCSE model.

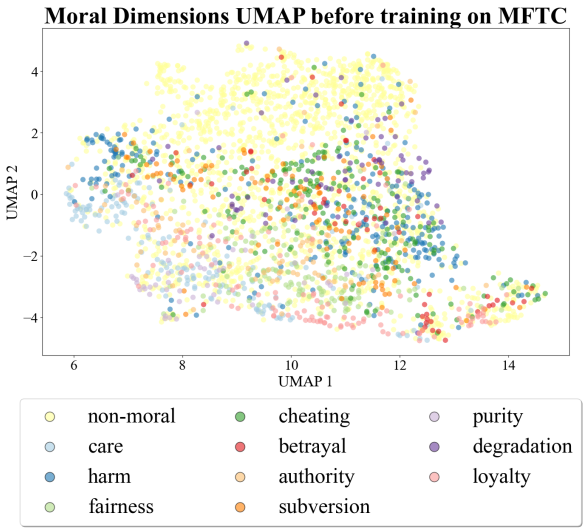

Figure B5: UMAP plot of Moral Foundations before training on the MFTC train set. Plotted using the MFTC test set.

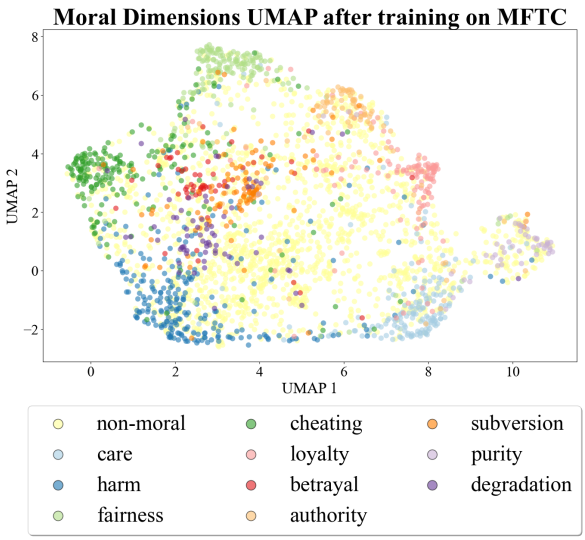

Figure B6: UMAP plot of Moral Foundations after training on the MFTC train set. Plotted using the MFTC test set.

B.7 Plotting with PCA

We plot with PCA using the MFTC train set to check whether a similar pattern can be found as in UMAP.

Figure B7 and Figure B8 show a similar pattern as in Figure B1 and Figure B2. The gap between the two clusters, vice and virtue, in the PCA plot gets wider after training with MFTC. The UMAP Plot, Figure B2, however, shows an even clearer separation.

Figure B9 and B10 also present the same facet as in the UMAP plots, Figure 1 and Figure 4. Training with MFTC obviously allows the model to learn the characteristics of MFT both in the PCA and the UMAP plots.

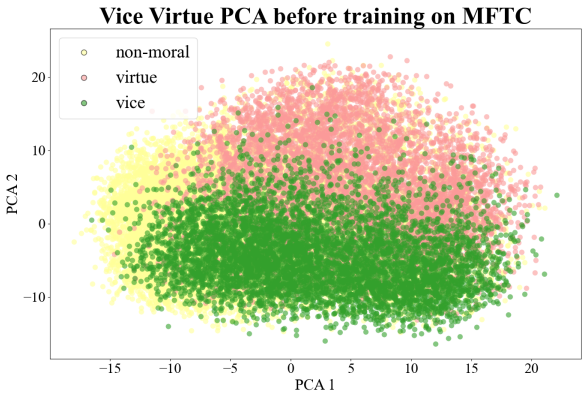

Figure B7: PCA plot of Vice and Virtue before training on MFTC (Virtue: Pink, Vice: Green)

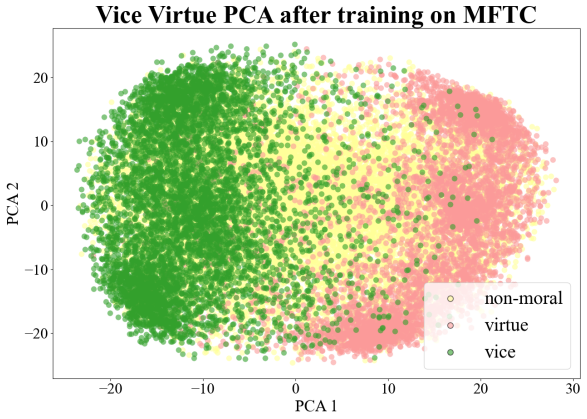

Figure B8: PCA plot of Vice and Virtue after training on MFTC (Virtue: Pink, Vice: Green)

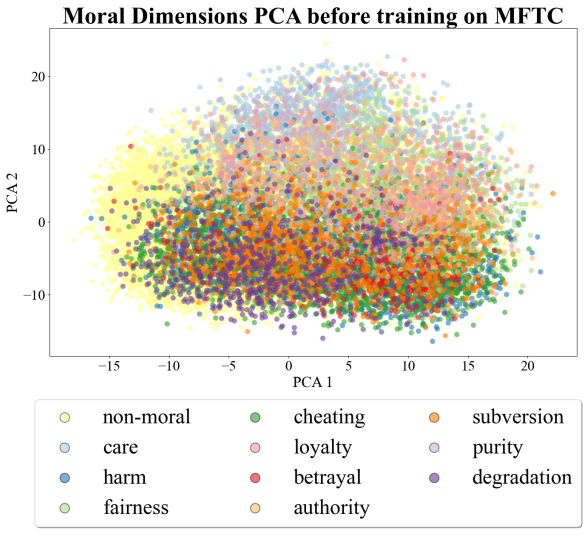

Figure B9: PCA plot of Moral Foundations before training on MFTC

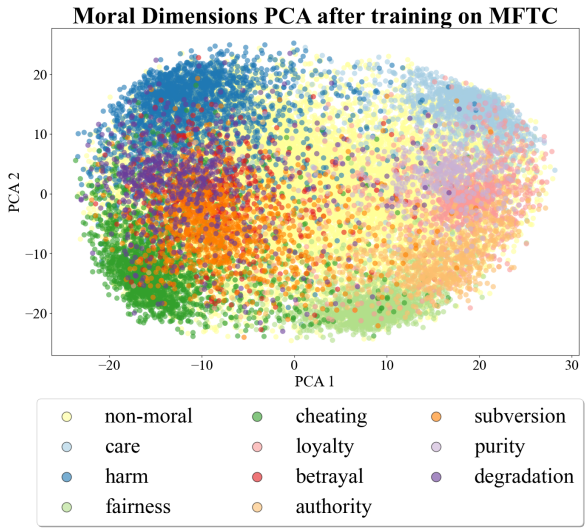

Figure B10: PCA plot of Moral Foundations after training on MFTC

B.8 Alignment and Uniformity

alignment and *unifomity* are the analysis metrics to assess the quality of the embedding space, which takes *alignment* between positive pairs and *uniformity* of the embedding space (Gao et al., 2021). They can be simply calculated with the following Equation 1 and Equation 2.

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \underset{(x,x^+)\sim p_{\text{pos}}}{\mathbb{E}} \left[\|f(x)-f(x^+)\|^2\right] \tag{1}$$

$$\mathcal{L}_{\text{uniform}}(f;t) \triangleq \log \underset{x,y \overset{\text{i.i.d}}{\sim} p_{\text{data}}}{\mathbb{E}} \left[e^{-2\|f(x)-f(y)\|^2}\right] \tag{2}$$

Table B10: *alignment* and *uniformity* Scores on MFTC classification dataset. For both, lower numbers are better

Model	alignment	uniformity
Sup.SimCSE	0.772	-2.27
Unsup. SimCSE	1.50	-3.12

Perfect *alignment* and *uniformity* are not our research goals. However, we computed the *alignment* and *uniformity* score using the test dataset, 10% of MFTC, to set a reference point of our embedding space. Table B10 displays the result of *alignment* and *uniformity* metrics. *Supervised* SimCSE outperforms in *alignment*, but gets a worse score in *uniformity* compared to *Unsupervised* SimCSE. The result is consistent with the findings in SimCSE (Gao et al., 2021) that *Supervised* SimCSE amends the *alignment* and *Unsupervised* SimCSE effectively improves *Uniformity*.