ME^2-BERT: Are Events and Emotions what you need for Moral Foundation Prediction?

Lorenzo Zangari[1], Candida M. Greco[1], Davide Picca[2], Andrea Tagarelli[1]

[1]DIMES, University of Calabria, Italy
[2]University of Lausanne, Switzerland
{lorenzo.zangari, candida.greco, tagarelli}@dimes.unical.it, davide.picca@unil.ch

Abstract

Moralities, emotions, and events are complex aspects of human cognition, which are often treated separately since capturing their combined effects is challenging, especially due to the lack of annotated data. Leveraging their interrelations hence becomes crucial for advancing the understanding of human moral behaviors. In this work, we propose ME^2-BERT, the first holistic framework for fine-tuning a pre-trained language model like BERT to the task of moral foundation prediction. ME^2-BERT integrates events and emotions for learning domain-invariant morality-relevant text representations. Our extensive experiments show that ME^2-BERT outperforms existing state-of-the-art methods for moral foundation prediction, with an average increase up to 35% in the out-of-domain scenario.

1 Introduction

Moral values embedded in language, reflecting what people think is "right" and "wrong", serve as a mirror for understanding human behaviors (Van de Poel and Royakkers, 2023). Morality is not a single, monolithic concept, but can be operationalized through multiple dimensions that capture the diversity of human moral reasoning (Schwartz, 1992). *Moral Foundations Theory* (MFT) builds on this premise by providing a foundational framework of five moral dimensions that are supposed to be widely accepted as universal to a large extent, i.e., MFT is subject to continous revisions (Atari et al., 2020). In recent years, the MFT has been instrumental in various studies to understand cultural differences, political ideologies, and the language used in moral discourses (Kobbe et al., 2020; Feinberg and Willer, 2015).

An emerging trend is the integration of the MFT with *Pre-trained Language Models* (PLMs), like BERT (Devlin et al., 2019), for exploring their abilities in moral foundation analysis tasks (Almeida et al., 2024). By learning lexical, semantic, and factual information from the training data, PLMs can also capture a range of cultural and moral biases embedded within it (Abdulhai et al., 2024). Since the linguistic choices of people reflect their moral values (Blankenship et al., 2021), PLMs can be effective in modeling moralities.

Moreover, the individuals' choices are also closely linked with the emotional states of people (Liu, 2020). *Emotions*, such as anger and joy, are complex states of feeling that result in physical and psychological reactions influencing both thought and behavior (Cambria et al., 2012). Emotions are closely linked with morality (Horne and Powell, 2016), indicating that leveraging emotional information can simplify morality analysis tasks.

Another crucial aspect is the representation of *events*. Acting as occurrences involving one or more entities, events can encapsulate a wide range of contextual information (e.g., actions, participants), which can be mapped to moral foundations. Thus, events might be helpful as a new layer of representation for textual documents, enabling their analysis from different perspectives, particularly for morality analysis tasks, where contextual information is crucial (Haidt and Graham, 2007).

In this work, we propose a PLM-based framework, namely ME^2-BERT, for **M**oral Foundation Prediction through **E**vents and **E**motions, through a fine-tuning of BERT. ME^2-BERT is specifically designed for leveraging emotions, events and moralities during the learning process, and can generalize across different types of data never seen during the training process. To this purpose, we devise a domain-adaptation strategy for learning robust *domain invariant representations* based on events, i.e., by aligning to a common embedding space both texts that contain events and texts that do not contain events. This is accomplished through a denoising auto-encoder, which is also constrained to learn emotion-aware text encodings through a

Proceedings of the 31st International Conference on Computational Linguistics, pages 9516–9532
January 19–24, 2025.

contrastive learning strategy. The learned embeddings are then fed to an adversarial classifier, for learning domain-invariant representations, and to a moral classifier for predicting moral foundations.[1]

ME^2-BERT is fine-tuned on the E2MoCase dataset (Greco et al., 2024)—specifically on its E2MoCase_full version—which, to the best of our knowledge, is the only available dataset linking moralities, emotions and events together within textual data (news articles). We argue that by forcing the model to learn domain-invariant representations from event-based domains, which both contain news from diverse media outlets, the model can effectively adapt to new datasets with different linguistic styles and biases, thus becoming a general moral foundation classifier. We summarize our contributions as follows:

1. We propose ME^2-BERT, a novel holistic framework for moral foundation prediction based on a fine-tuning of PLM like BERT by integrating moralities, emotions, and events.

2. We define a domain identification strategy based on events, and an emotion-aware denoising auto-encoder module, which acts as adversary of an event-based domain classifier for learning domain-invariant representations.

3. Our experimental evaluation on existing datasets with moral foundation annotations has shown the significance of our framework against several methods for moral foundation prediction, including lexicon-based, BERT-based and LLM approaches, with average percentage increase in F1-score ranging from 15% to 33% in average. Also, the experimental results indicate the usefulness of all components of ME^2-BERT, including those exploiting emotions and events.

We provide the source code and trained model at https://mlnteam-unical.github.io/resources/.

2 Preliminaries

In this section, we introduce preliminary definitions. A table of the notations used throughout the paper is provided in the Appendix.

[1]Note that we follow the established practices to refer to high-resource languages, particularly English, while acknowledging the inherent risk of cultural biases.

Moral Foundation Theory. Our work is grounded in the Moral Foundations Theory (MFT) (Haidt and Joseph, 2004), which provides a theoretical framework for operationalizing the concept of human morality. It assumes that five distinct *dimensions or foundations*, each consisting of a *duality of vice and virtue*, can describe all moral dilemmas: *Care (Cr.) / Harm (Hr.)* focusing on empathy and protection versus infliction of suffering; *Fairness (Fr.) / Cheating (Ch.)*, centered on upholding justice and integrity versus deceit and exploitation; *Loyalty (Ly.) / Betrayal (Br.)*, promoting allegiance to one's group versus acts of betrayal; *Authority (Au.) / Subversion (Sb.)*, valuing obedience to societal norms and traditions versus challenges to authority; *Purity (Pr.) / Degradation (Dr.)*, emphasizing the sanctity of what is considered sacred versus its defilement.

Problem definition: Moral Foundation Prediction. We define moral foundation prediction as the task of predicting the moral foundations contained in textual data according to the principles outlined by the MFT. We are given a dataset $\mathcal{D} = \{(d_i, \mathbf{y}_i)\}$, where d_i represents a text and $\mathbf{y}_i \in \mathbb{R}_*^{|\mathcal{M}|}$ is a vector of non-negative real-value scores associated with the dimensions of the morality theory, i.e., in our setting MFT with $|\mathcal{M}| = 5$. Our goal is to learn a function $f : \mathcal{T} \rightarrow \{0, 1\}^{|\mathcal{M}|}$ that predicts the moral foundations for any $d_i \in \mathcal{T}$, where $\mathcal{T}$ is the space of all possible input texts. Whenever none of the MFT dimensions are present in a text d_i, its content is treated as *non-moral (Nm.)* (Hoover et al., 2020a). Following other works (Trager et al., 2022), we frame the problem as a multi-label classification task. This is in line with the fact that most of the existing moral data are provided with binary labels.

Domain adaptation. *Unsupervised Domain adaptation (UDA)* is a transfer learning technique designed to address domain shifts in data distribution, enhancing out-of-domain prediction performance (Ben-David et al., 2010). UDA involves unsupervised learning as it relies on labeled data from a source domain while assuming no labels for a target domain. Formally, let $\mathcal{D}_s = \{(d_i^{(s)}, \mathbf{y}_i^{(s)})\}_{i=1}^{n_s}$ and $\mathcal{D}_t = \{d_j^{(t)}\}_{j=1}^{n_t}$ denote the labeled source domain and unlabeled target domains, respectively, where n_s and n_t are the number of source and target samples and $n = n_s + n_t$. Due to the domain shift, both the marginal distribution ($\mathbb{P}$) and the conditional distributions ($\mathbb{Q}$) of the two domains

differ, i.e., $\mathbb{P}_s(d^{(s)}) \neq \mathbb{P}_t(d^{(t)})$, $\mathbb{Q}_s(y^{(s)}|d^{(s)}) \neq \mathbb{Q}_t(y^{(t)}|d^{(t)})$, and are not known a priori. The goal of UDA for moral foundation prediction is to learn a function $f : \mathcal{T} \rightarrow \mathbb{R}^{|\mathcal{M}|}$ that predicts the moralities of samples from the target domain by minimizing domain discrepancy between source and target (Singhal et al., 2023).

Common domain adaptation methods include *domain adversarial networks*, which introduce a domain discriminator to enforce the learning of domain-invariant embeddings (Ganin et al., 2016). This approach has been effective in predicting moral foundations from text (Guo et al., 2023). Another strategy, based on the idea that a domain-adaptive framework should effectively reconstruct target domain data, involves learning domain-invariant embeddings using a *denoising auto-encoder* to reconstruct the original embeddings from the corrupted ones (Ghifary et al., 2016).

Why including events and emotions for moral foundation prediction? When addressing moral foundation analysis tasks, relying solely on textual information may be insufficient, as morality involves a wide array of human behaviors, including emotions and situational contexts (Haidt and Joseph, 2004). Moral principles are deeply embedded in the events people experience. Events, defined as specific occurrences at a particular time and place involving one or more participants (Xiang and Wang, 2019), provide contextual information, often marked by trigger verbs or nouns. Events do not merely provide background; they actively contribute to the moral interpretation of situations. They also serve as powerful tools for data augmentation, which is crucial in the moral domain, due to the scarcity of reliable, annotated data (Kobbe et al., 2020). By incorporating events, we can diversify the input without needing multiple data sources, thus allowing models to better generalize and learn domain-invariant representations. At the same time, the link between emotions and morality is well-established, though the exact nature of this relationship is still debated (Cameron et al., 2015).

By jointly integrating events and emotions for moral foundation prediction, we gain a twofold advantage. First, the use of events within texts creates a domain shift in data distribution, allowing the framework to learn domain-invariant representations that enhance its ability to generalize across different moral domains. Second, by leveraging emotions, the model can better capture the subjective nature of morality due to the intrinsic link between these two aspects (Tekin and Ekici, 2023). Figure 1 shows an example of text associated with emotions, moralities and events.

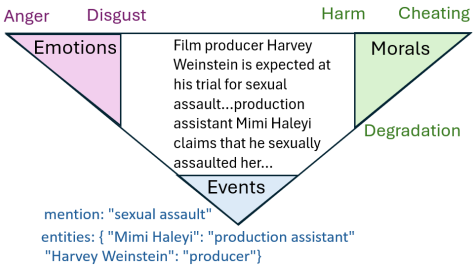


Figure 1: Text segment with emotions, moralities and events extracted from E2MoCase (Greco et al., 2024).

3 Related works

Early works for moral foundation prediction are lexicon-based, utilizing lists of words linked to moral foundations. The Moral Foundations Dictionary (MFD) (Graham et al., 2009) is one of the first lexicons developed for this purpose. Its extensions, MFDv2 (Frimer, 2019) and the extended Moral Foundation Dictionary (eMFD) (Hopp et al., 2021), expand the MFD by including more words and a larger set of annotated text samples. Distributed Dictionary Representations (DDR) (Garten et al., 2018) combines the MFD with word embeddings, representing a concept in the semantic space through the vector representation of the words in the MFD. MoralStrength (Araque et al., 2020) and LibertyMFD (Araque et al., 2022) further enhance the MFD by quantifying the relevance and strength of words related to the five moral foundations. These methods rely on predefined word lists and lack adaptability to diverse linguistic contexts.

Recent works have collected datasets annotated with moral foundations and used them for training deep-learning models. Hoover et al. (2020a) introduced the Moral Foundation Twitter Corpus (MFTC), which comprises seven distinct Twitter datasets focused on morality relevant issues, whose validity was tested using a LSTM model. In a similar vein, Trager et al. (2022) developed the Moral Foundation Reddit Corpus (MFRC) and established a series of baseline models employing BERT. Liscio et al. (2022) approached each dataset within the MFTC as a separate domain and trained BERT under different configurations.

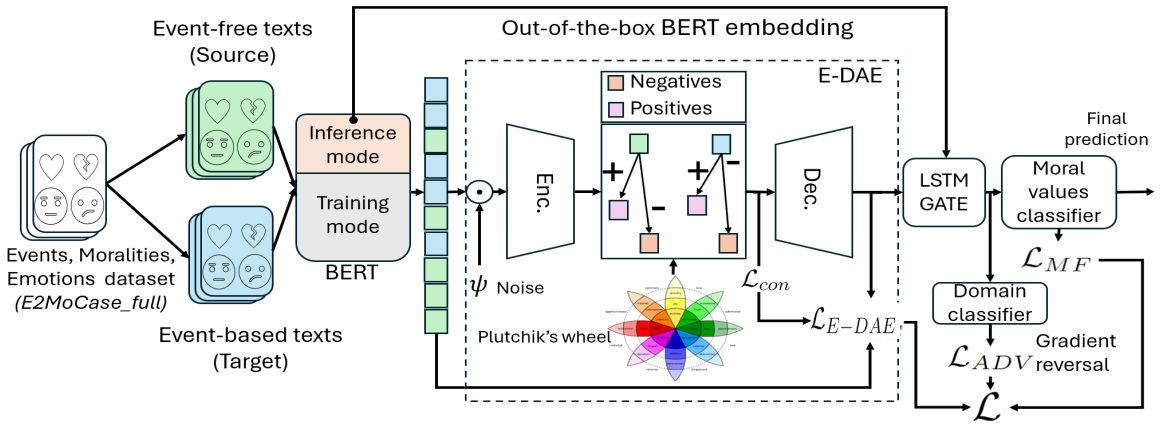


Figure 2: The ME^2-BERT framework for moral foundation prediction. ME^2-BERT leverages emotion and event information for learning domain-invariant morality-relevant text representations.

DAMF (Guo et al., 2023) and MoralBERT (Preniqi et al., 2024) are approaches relatively close to our work. In fact, both employ domain adaptation to train BERT on heterogeneous data sources; however, DAMF was trained in a semi-supervised fashion, having access to some data from the target domain, while MoralBERT showed a significant drop in performance in out-of-domain scenarios, indicating that it cannot effectively handle the domain shift problem between source and target data. Zhang et al. (2024) released the MoralEvent dataset, comprising of news article annotated with events and moralities, while Greco et al. (2024) connected moralities, events and emotions.

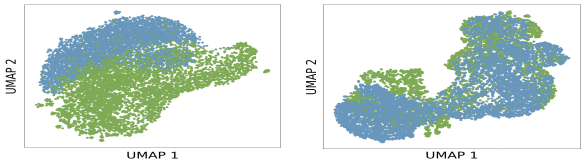
Like DAMF and MoralBERT, we employ a domain adaptation strategy to build a model for morality inference on unseen data. However, we leverage events and emotions to build a more robust domain-invariant representation. While events, often treated as graph-structured data, are mainly used for text augmentation (Shorten et al., 2021), and emotions are closely linked to morality (Tekin and Ekici, 2023; Ugazio et al., 2012), their combined use in deep learning for moral foundation prediction has remained unexplored so far. Their joint use enables training a domain-adaptive model driven by emotional information while relying on a single data source rather than depending on several heterogeneous datasets.

4 Methodology

The overall architecture of ME^2-BERT is shown in Fig. 2. ME^2-BERT is a framework for fine-tuning PLMs incorporating (i) events to learn domain-invariant text embeddings, and (ii) emotions to drive the learning of morality-relevant text embeddings for the moral foundation prediction task. Firstly, source and target domains are selected based on whether they do not contain events or contain events, respectively. In our training dataset, i.e., E2MoCase, this events-based selection ensures a strong domain shift between the two types of corpora. Then, a BERT-based model generates the embeddings for source and target texts, which are give in input into a *denoising auto-encoder* (DAE) (Vincent et al., 2008). The DAE is trained to reconstruct the BERT embeddings while being informed with emotional information using a contrastive learning strategy. The reconstructed embeddings are fed into both an adversarial-learning-based domain classifier (Ganin et al., 2016) to map the embeddings from different domains into a common space, and a moral foundation classifier.

4.1 Event-based domain identification

By exploiting the fine-tuning of ME^2-BERT on the E2MoCase dataset, we use events to create a domain shift in features and labels distribution between text segments without events (referred to as source domain) and text segments with events (referred to as target domain), thus forcing the model to learn robust, domain-invariant representations. When events are available, we model a text segment as a JSON object representing each event as a tuple of its corresponding trigger words and involved entities (cf. Fig. 9 in Appendix B). Indeed, it is common for a narrative to be introduced through events, e.g., to mold public opinion (Zhang et al., 2021). Also, our distinction of the source and target texts is useful to make the model align event-free and event-based text representations, enhancing its generalization capabilities.



(a) Out-of-the-box BERT embedding.

(b) Embedding learned by ME^2-BERT.

Figure 3: UMAP embedding. Green and blue points show the [CLS] token encoding of the source and target domains, respectively.

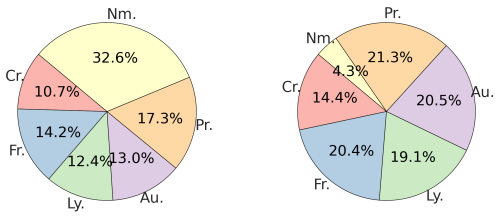


Figure 4: Distribution of the moral labels for source domain (on the left) and target domain (on the right).

Our domain identification strategy could not always result in a domain shift, and in general, it is required to analyze the fine-tuning data (cf. Sect. 8). Figure 3a shows the UMAP plot (McInnes et al., 2018) of the BERT embeddings of the source and target domains on E2MoCase, which differ in the embedding space, while Fig. 4 shows the labels distribution, revealing that they have also different moral focus. The predominance of non-moral data in the source domain suggests that event-free paragraphs often carry "neutral" content from the morality viewpoint. This can lead to data imbalance issues (Liscio et al., 2022).

4.2 Integration of emotional values

We provide emotional information through the Plutchik's wheel of emotions (Plutchik, 2001), which defines eight basic emotions arranged in pairs of opposites. We group the primary emotions of the Plutchik's wheel into opposing pairs, resulting in four broad categories: *anger/fear*, *trust/disgust*, *joy/sadness*, and *surprise/anticipation*. These categories are used within our contrastive learning framework to guide the model's training to incorporate emotional state information in the prediction of moral foundations.

4.3 Emotion-aware Denoising Auto-Encoder

The Emotion-aware Denoising Auto-Encoder (E-DAE) module in our ME^2-BERT is designed to learn robust embeddings leveraging emotion information for both event-based and event-free texts (i.e., domain). By utilizing a denoising auto-encoder as a transformation function, we filter out domain-specific noise and features contained in the BERT-based encodings (Guo et al., 2023). This approach improves robustness and generalizability so as to better support the alignment of embeddings across different domains (Wang et al., 2021; Lopez-Avila and Suárez-Paniagua, 2023). Following Clinchant et al. (2016), we apply the E-DAE to the union of the source and target samples. Given d_i, let $\mathbf{x}_i = BERT(d_i)$ be its BERT embedding generated during training. We corrupt $\mathbf{x}_i$ with noise ψ, i.e., $\mathbf{x}'_i = \mathbf{x}_i \odot \psi$, where ψ is a random variable sampled from a Bernoulli distribution with probability p, and $\odot$ indicates element-wise multiplication. The auto-encoder aims to reconstruct the BERT embeddings without noise ($\mathbf{x}_i$) given the corrupted embeddings $\mathbf{x}'_i$:

$$\mathbf{h}_i = f_{enc}(\mathbf{x}'_i), \quad \hat{\mathbf{x}}_i = f_{dec}(\mathbf{h}_i), \tag{1}$$

where $\hat{\mathbf{x}}_i$ is the reconstructed encoding from the auto-encoder; f_{enc} and f_{dec} are the encoder and decoder functions, corresponding to two-layer MLPs; $\mathbf{h}_i$ is the bottleneck representation learned by the encoder. The reconstruction loss is given by the Mean Squared Error (MSE), shown in Eq. 2:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2, \tag{2}$$

where n is the number of text segments. To favor the discovery of latent representations that are emotion-aware, we add a separate contrastive term for each domain that encourages samples with the same emotions to be closer in the embedding space, while those with different emotions are pushed far apart. This contrastive objective ensures that the learned representations capture emotion information, which is crucial for morality analysis tasks (Tekin and Ekici, 2023). Eq. 3 defines the source loss $\mathcal{L}_{\text{tr}}^{(s)}$, which is a triplet margin objective:

$$\mathcal{L}_{\text{tr}}^{(s)}(i,p,r) = \max\Big(0, c(\mathbf{h}_i,\mathbf{h}_p) - c(\mathbf{h}_i,\mathbf{h}_r) + m\Big), \tag{3}$$

where $\mathbf{h}_p$ and $\mathbf{h}_r$ are the latent representations of the positive and negative samples associated with the text segment d_i from the source domain; m is the margin enforced between positive and negative pairs and $c(\cdot,\cdot)$ is the cosine similarity distance. Analogous contrastive term is used for the samples belonging to the target domain ($\mathcal{L}_{\text{tr}}^{(t)}$). By applying a separate contrastive term to the source and

target domains we ensure that important emotional information are preserved, thus preventing the loss of domain-specific information which might be blurred during the denoising process (Kang et al., 2019; Lopez-Avila and Suárez-Paniagua, 2023). The pairs of positive and negative samples are selected based on the Plutchik's categories described in Sect. 4.2, where each text instance is associated with the emotions having the highest score in E2MoCase. The overall loss function for training the E-DAE corresponds to the sum of reconstruction loss and the contrastive loss:

$$\mathcal{L}_{E-DAE} = \mathcal{L}_{MSE} + \mathcal{L}_{con}, \tag{4}$$

where $\mathcal{L}_{con} = \mathcal{L}_{\text{tr}}^{(s)} + \mathcal{L}_{\text{tr}}^{(t)}$, and $\mathcal{L}_{\text{tr}}^{(s)}$, $\mathcal{L}_{\text{tr}}^{(t)}$ are the source and target contrastive terms, respectively.

4.4 LSTM gate mechanism

The reconstruction provided by the E-DAE module may lead to the loss of semantic information inherent in the original text. To overcome this issue, we propose a LSTM-style gated fusion mechanism that integrates general pre-trained knowledge with the fine-tuned representations. Let $\mathbf{x}_i^{(oob)}$ denote the embedding produced by the out-of-the-box BERT model, and $\hat{\mathbf{x}}_i$ the embeddings produced by E-DAE. The gating mechanism is defined as:

$$\begin{aligned} &\mathbf{f}_i = \sigma(\mathbf{W}_f \mathbf{x}_i^{(oob)} + \mathbf{b}_f), \quad \mathbf{q}_i = \sigma(\mathbf{W}_q \hat{\mathbf{x}}_i + \mathbf{b}_q), \\ &\mathbf{c}_i = \mathbf{f}_i \odot \mathbf{x}_i^{(oob)} + \mathbf{q}_i \odot \hat{\mathbf{x}}_i, \quad \mathbf{o}_i = \sigma(\mathbf{W}_o \mathbf{c}_i + \mathbf{b}_o), \\ &\tilde{\mathbf{x}}_i = \mathbf{o}_i \odot \tanh(\mathbf{c}_i), \end{aligned} \tag{5}$$

where σ is the sigmoid function; $\mathbf{W}_f, \mathbf{W}_q, \mathbf{W}_o$ are weight matrices; $\mathbf{b}_f, \mathbf{b}_q, \mathbf{b}_o$ are bias vectors; $\odot$ denotes element-wise multiplication and $\mathbf{f}_i$, $\mathbf{q}_i$ and $\mathbf{o}_i$ are the forget, input and output gates, respectively. $\mathbf{f}_i$ and $\mathbf{q}_i$ determine the amount of information to retain from $\mathbf{x}_i^{(oob)}$, and $\hat{\mathbf{x}}_i$, resp., while $\mathbf{c}_i$ is the vector state combining the out-of-the-box and the reconstructed embeddings; $\mathbf{o}_i$ controls the contribution of $\mathbf{c}_i$ to the final gated representation $\tilde{\mathbf{x}}_i$.

4.5 Event classifier

To strengthen the domain invariance effect, we integrate an adversarial learning module applied to the gated representations, i.e., $\tilde{\mathbf{x}}_i$. By introducing an adversarial discriminator that challenges the model to produce embeddings indistinguishable across domains—thus acting as adversary of the E-DAE module— we further enforce the learning of domain-invariant embeddings (Ganin et al., 2016; Guo et al., 2023). The loss function for the domain classifier is defined in Eq. 6, which is to be maximized during the training through a gradient reversal layer (Ganin et al., 2016).

$$\begin{aligned} \mathcal{L}_{ADV} = -\frac{1}{n}\sum_{i=1}^{n} \Big[& y_i^{(D)} \log(\sigma(g(\tilde{\mathbf{x}}_i))) + \\ & + (1 - y_i^{(D)}) \log(1 - \sigma(g(\tilde{\mathbf{x}}_i))) \Big], \end{aligned} \tag{6}$$

where g is the domain classifier (MLP), which aims to distinguish between the two domains; σ is the sigmoid function, and $y_i^{(D)}$ is the ground truth label indicating which domain the input text belongs to.

4.6 Moral foundation classifier

To predict the moral foundations, we use a MLP shared for both domains, which is fed with the denoised and domain-invariant BERT embedding $\tilde{x_i}$. Since we address the problem as a multi-label classification task, the loss function, i.e., $\mathcal{L}_{MF}$, is a Cross-Entropy with sigmoid activation. ME^2-BERT is trained to optimize the following loss:

$$\mathcal{L} = \mathcal{L}_{E-DAE} + \mathcal{L}_{MF} - \lambda_{dom}\mathcal{L}_{ADV}, \tag{7}$$

where λ_{dom} is an hyperparameter indicating the importance of the adversarial learning module.

5 Computational complexity aspects

We discuss the time complexity of ME^2-BERT, under the following assumptions: the sequence length of each sample in the batch is set to the maximum value T, and the hidden dimension of each neural model out of the BERT encoder is d_b.

Generating the BERT embeddings requires $\mathcal{O}(n_b L T^2 d_B)$ (Vaswani et al., 2017), where L is the number of BERT layers, n_b is the number of samples in the batch, and d_B is the hidden dimension. With a bottleneck layer of size d_b, and using two-layer MLPs for both the encoder and decoder, the autoencoder requires $\mathcal{O}(n_b d_B d_b + n_b d_b^2) \approx \mathcal{O}(n_b d_B d_b)$, since $d_b \ll d_B$. The LSTM gate requires $\mathcal{O}(n_b d_B^2)$ operations, stemming from the linear transformations and the additional element-wise operations described in Eq. 5. The moral classifier requires $\mathcal{O}(d_B d_b)$ operations. Therefore, during inference, the cost is $\mathcal{O}(n_b L T^2 d_B + n_b d_B^2)$.

At training time, the domain classifier adds a cost of $\mathcal{O}(d_B d_b)$ and we need to compute the loss functions. The triplet margin loss is the most computationally expensive, requiring cosine distance calculations for up to $\mathcal{O}(n_b^2)$ triplets.

6 Experimental evaluation

Evaluation goals. We design our experimental evaluation to pursue the following objectives: (i) to measure the effectiveness of ME^2-BERT w.r.t. SOTA NLP methods and LLMs for moral foundation prediction; (ii) to carry out an ablation analysis which shows the impact of each component of ME^2-BERT; (iii) to validate our event-based domain identification strategy; (iv) to evaluate the performance of ME^2-BERT in the single-label setting, where each moral foundation is treated independently, and (v) to analyze the abilities of ME^2-BERT to detect the polarity of each moral foundation.

Datasets. As previously said, we fine-tuned ME^2-BERT on E2MoCase, specifically the E2MoCase_full dataset. We evaluate all methods on three datasets that are manually annotated with moralities: Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020b), Moral Foundation Reddit Corpus (MFRC) (Trager et al., 2022), and the Extended MFD (eMFD) (Hopp et al., 2021).

Competing methods. We compare ME^2-BERT with different classes of competitors: (i) *BERT-based methods* specifically trained for moral foundation prediction, including BERT-base trained on E2MoCase dataset (hereinafter BERT-E2MoCase), the baselines proposed by Trager et al. (2022), namely BERT trained on MFTC (*BERT-MFTC*) and BERT trained on MFRC (*BERT-MFRC*). Additionally, we include BERT trained on both MFTC and MFRC (BERT-MFTRC), *DAMF* and *MoralBERT*. (ii) *LLMs* with zero-shot inference strategy, which include three of the most recent open models: *Llama-3.1*, *Gemma-2* and *Mistral-Nemo*; (iii) *lexicon-based methods* designed for moral classification, including *MoralStrength* and *DDR*.

Experimental setting. We evaluate the performance of each model in the out-of-domain setting, where training and test data come from different domains. Although ME^2-BERT is independent of the PLM used, we employ BERT base uncased to be fair with our competitors—*although, we also experimented with a selection of Sentence-Transformers* (cf. Appendix F). ME^2-BERT and DAMF are fine-tuned on the E2MoCase dataset, while all the other BERT-base models are fine-tuned on the MFRC and/or the MFTC datasets (e.g., BERT-MFRC). All BERT-based models are evaluated utilizing a 5-fold cross validation strategy, except for MoralBERT for which we use its single-label version in inference mode, as provided by the authors. Since it uses the the same strategy as DAMF, and it is fine-tuned on MFRC and MFTC datasets, we report its performance solely on eMFD data to maintain consistency with the out-of-domain setting. For LLMs and lexicons, we report the average scores over five runs. More details are reported in the Appendix E.

	Cr.	Fr.	Ly.	Au.	Pr.	Nm.	AVG
MFRC							
ME^2-BERT	**0.636**	**0.585**	0.345	**0.490**	**0.363**	**0.669**	**0.515**
DAMF	0.457	0.535	0.260	0.419	0.311	0.520	0.417
BERT-MFTC	0.619	0.475	0.296	0.321	0.250	0.507	0.411
BERT-E2MoCase	0.481	0.495	0.347	0.396	0.314	0.610	0.440
Llama-3.1	0.481	0.481	0.298	0.392	0.273	0.010	0.323
Mistral-Nemo	0.496	0.505	**0.353**	0.405	0.233	0.245	0.373
Gemma-2	0.519	0.484	0.312	0.394	0.252	0.103	0.344
MoralStrength	0.434	0.463	0.209	0.335	0.183	0.120	0.291
DDR	0.465	0.276	0.201	0.377	0.280	0.623	0.371
MFTC							
ME^2-BERT	**0.688**	**0.673**	**0.551**	0.521	**0.453**	**0.546**	**0.572**
DAMF	0.540	0.583	0.486	0.485	0.397	0.459	0.492
BERT-MFRC	0.568	0.499	0.325	0.435	0.342	0.527	0.449
BERT-E2MoCase	0.544	0.581	0.432	0.380	0.314	0.510	0.460
Llama-3.1	0.627	0.609	0.469	**0.541**	0.452	0.144	0.474
Mistral-Nemo	0.631	0.518	0.459	0.483	0.416	0.407	0.486
Gemma-2	0.642	0.603	0.466	0.524	0.419	0.271	0.487
MoralStrength	0.589	0.531	0.501	0.494	0.379	0.442	0.489
DDR	0.503	0.392	0.314	0.375	0.412	0.432	0.405
eMFD							
ME^2-BERT	0.309	0.225	0.217	0.227	**0.220**	0.373	**0.262**
DAMF	0.283	0.231	0.192	0.256	0.185	0.317	0.244
MoralBERT	0.279	0.226	**0.220**	0.235	0.111	0.007	0.196
BERT-MFTRC	0.296	0.229	0.210	0.236	0.064	0.422	0.243
BERT-E2MoCase	0.297	0.161	0.101	0.155	0.037	0.480	0.210
Llama-3.1	0.330	0.287	0.145	0.273	0.201	0.207	0.240
Mistral-Nemo	0.314	0.267	0.170	0.270	0.167	0.371	0.260
Gemma-2	**0.333**	**0.293**	0.209	**0.279**	0.178	0.211	0.250
MoralStrength	0.244	0.220	0.208	0.200	0.182	0.436	0.248
DDR	0.210	0.118	0.153	0.191	0.080	**0.484**	0.206

Table 1: Comparative evaluation: F1-scores. Best results are in bold, second-best results are underlined.

6.1 Results

Comparative evaluation. Table 1 reports the F1-score values achieved by ME^2-BERT and the competing methods on each dataset. ME^2-BERT stands out as the best approach yielding an average F1-score of 0.450 across all datasets, while the second best method is DAMF (0.384), which also utilizes a domain adaptation strategy, but it does not leverage events and emotions. Despite the LLMs benefit from extensive pre-training on diverse text sources, the best LLM (Mistral-Nemo) yields low performance than ME^2-BERT, especially on the MFTC and MFRC datasets (with percentage decrease of about 20% in average). At the moral foundation level, ME^2-BERT consistently performs well across all moral foundations, particulary for the MFRC and MFTC datasets. On the eMFD dataset, ME^2-BERT is outperformed by

	Cr.	Fr.	Ly.	Au.	Pr.	Nm.	AVG
MFRC							
BERT-E2MoCase	0.481	0.495	**0.347**	0.396	0.314	0.610	0.440
ME^2-BERT w/o A.	0.566	0.555	0.294	0.466	0.327	**0.672**	0.480
ME^2-BERT w/o C.	0.587	0.541	0.278	0.451	0.316	0.636	0.468
ME^2-BERT w/o A-C.	0.583	0.525	0.304	0.421	0.328	0.623	0.464
ME^2-BERT w/o G.	0.629	0.577	0.332	0.481	0.360	**0.672**	0.510
ME^2-BERT	**0.636**	**0.585**	0.345	**0.490**	**0.363**	0.669	**0.515**
MFTC							
BERT-E2MoCase	0.544	0.581	0.432	0.380	0.314	0.510	0.460
ME^2-BERT w/o A.	0.624	0.623	0.525	0.457	0.419	0.509	0.526
ME^2-BERT w/o C.	0.650	0.622	0.528	0.478	0.422	0.517	0.536
ME^2-BERT w/o A-C.	0.615	0.610	0.501	0.518	0.392	0.503	0.523
ME^2-BERT w/o G.	0.687	**0.682**	0.529	0.514	0.431	0.519	0.560
ME^2-BERT	**0.688**	0.673	**0.551**	**0.521**	**0.453**	**0.546**	**0.572**
eMFD							
BERT-E2MoCase	0.298	0.162	0.101	0.155	0.037	**0.480**	0.205
ME^2-BERT w/o A.	0.284	0.246	0.233	**0.239**	0.206	0.297	0.251
ME^2-BERT w/o C.	0.300	0.193	0.179	0.157	0.134	0.427	0.232
ME^2-BERT w/o A-C.	0.284	**0.250**	**0.242**	0.218	**0.228**	0.130	0.225
ME^2-BERT w/o G.	0.298	0.218	0.221	**0.239**	0.199	0.285	0.243
ME^2-BERT	**0.309**	0.225	0.217	0.227	0.220	0.373	**0.262**

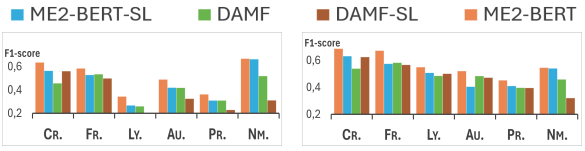
Table 2: Ablation study: F1-scores. Best results are in bold, second-best results are underlined.

	Cr.	Fr.	Ly.	Au.	Pr.	Nm.	AVG
MFRC							
ME^2-BERT	**0.387**	**0.226**	**0.139**	**0.162**	**0.144**	0.283	**0.224**
BERT-ME	0.348	0.006	0.000	0.003	0.000	**0.636**	0.165
MFTC							
ME^2-BERT	**0.502**	**0.162**	**0.132**	**0.262**	**0.206**	0.161	**0.238**
BERT-ME	0.360	0.011	0.000	0.008	0.000	**0.376**	0.126
eMFD							
ME^2-BERT	**0.210**	**0.219**	**0.085**	**0.170**	**0.089**	0.240	**0.169**
BERT-ME	0.135	0.006	0.000	0.014	0.000	**0.459**	0.102

Table 3: F1-scores of ME^2-BERT and BERT-base (BERT-ME) fine-tuned on MoralEvent. Best results are in bold.

LLMs in 3 out of 5 moral categories. Nevertheless, ME^2-BERT remains competitive, achieving performance comparable to LLMs on average.

Ablation analysis. Table 2 summarizes the F1-scores by ME^2-BERT and several simplified variants (cf. abbreviations' descriptions in Appendix, Table 5). We observe that the best results are achieved by the full ME^2-BERT, while BERT-E2MoCase is the worst model, demonstrating the effectiveness of the domain-adaptive approach. As emotions and/or events are integrated, the performance tends to increase. The variant with only the DAE module, excluding the adversarial learning and contrastive modules (dubbed ME^2-BERT *w/o A-C.*), shows an average performance increase of nearly 10% over BERT-E2MoCase. Emotions also improve performance, as indicated by the scores achieved by the variant not using the contrastive learning module (dubbed ME^2-BERT *w/o C.*). The gate component (dubbed ME^2-BERT *w/o G.*) also yields benefits.



(a) Performance on MFRC (b) Performance on MFTC

Figure 5: ME^2-BERT and DAMF performance in the single-label (-SL) vs multi-label classification settings.

Fine-tuning on MoralEvent. To further assess the validity of our event-based domain identification strategy, we fine-tune ME^2-BERT and BERT-base on the *MoralEvent* dataset (Zhang et al., 2024), which contains about 400 news article labeled with morals and events. Since emotion labels are not provided, we use the version of ME^2-BERT w/o contrastive learning (ME^2-BERT w/o C.). Table 3 reports the F1-scores achieved by the two methods. The overall performances are lower than the version trained on E2MoCase, likely due to the smaller size of MoralEvent. However, ME^2-BERT yields significantly higher F1-scores on all moral foundations, while BERT fails to identify them (both precision and recall are zero for several dimensions), resulting in a high score for the non-moral label.

Single-label classification. We explore the moral foundation prediction problem in the single-label setting, where a single model is fine-tuned for each moral foundation. Figure 5 shows the performance of ME^2-BERT and the best competing method, i.e., DAMF, on the MFRC and MFTC datasets. Results for eMFD are shown in the Appendix. Interestingly, both ME^2-BERT and DAMF tend to perform better in the multi-label setting than their single-label counterparts (ME^2-BERT-SL and DAMF-SL, respectively). While previous works (Liscio et al., 2022; Trager et al., 2022) found that the multi-label setting is more challenging, we argue that this behavior may stem from their domain-adaptive nature, enabling them to better handle cross-dimension dependencies.

Moral polarity detection. Figure 6 shows the performance of ME^2-BERT in predicting the polarity (virtue/vice) of each moral foundation, such as care/harm, on the MFTC dataset, compared to the best competing methods for each category (e.g., Llama-3.1 for LLMs). The scores achieved on the eMFD dataset are shown in the Appendix F. Overall, ME^2-BERT outperforms the competitors on most moral foundations, yielding the highest av-

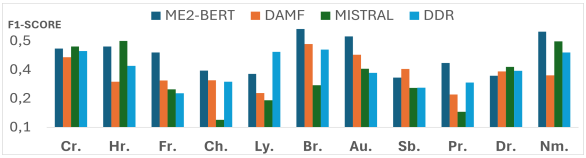
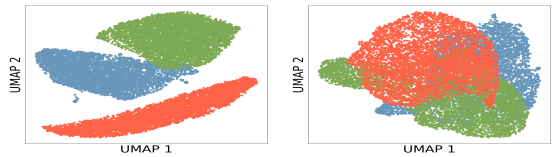


Figure 6: Prediction of virtue/vice moral foundations.



(a) Out-of-the-box BERT embedding (b) Embeddings learned by ME^2-BERT

Figure 7: UMAP of the dataset embeddings. Green, blue and orange points indicate samples from the MFRC, MFTC and eMFD datasets, respectively.

erage F1-score (0.427), followed by DDR (0.353) and DAMF (0.334).

Visual interpretation. Figure 3b shows the UMAP of the embeddings learned by ME^2-BERT on E2MoCase, which indicates effective domain-alignment despite the initial feature distribution differences. Considering the evaluation data, Fig. 7 shows their UMAP visualization using BERT (a) and ME^2-BERT embeddings (b), respectively. The out-of-the-box BERT embeddings are separately clustered, reflecting distinct domain-specific features. By contrast, ME^2-BERT aligns these datasets into a more overlapping space, as it had learned to capture high-level, shared textual features rather than domain-specific ones. This would indicate that ME^2-BERT can abstract moral concepts that generalize across different domains, improving its ability in predicting moral foundations.

Qualitative analysis of the responses. Figure 8 shows the ME^2-BERT prediction outputs on some instances of E2MoCase. The two plots on top correspond to examples of successful prediction of all moral foundations associated with the text segments, aligning with the nature of the events and the societal challenges they reflect. In particular, Fig. 8a shows how ME^2-BERT is able to capture moral foundations in the context of a murder case, while Fig. 8b corresponds to Authority/Subversion and Purity/Degradation in a narrative about social protests. In Fig. 8c, the model misses the Loyalty/Betrayal foundation, likely due to the focus of the text on violence, harm and justice, which aligns more with Care/Harm and Fairness/Cheating and

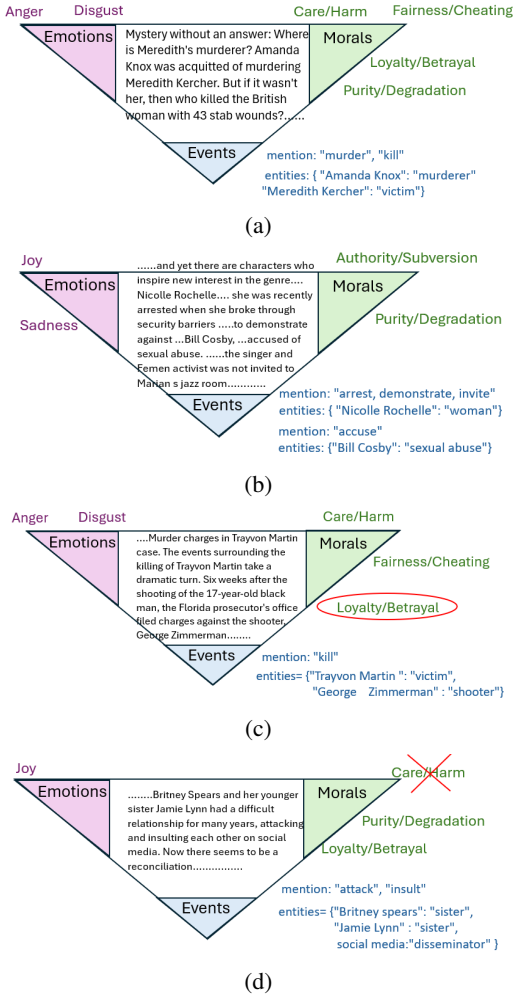


Figure 8: Examples of moral foundation prediction by ME^2-BERT. The red circle and red cross indicate false negatives and false positives, respectively.

are further amplified by the emotions of anger and disgust (Haidt et al., 2003; Graham et al., 2013). In Fig. 8d, where the text is about interpersonal conflict and reconciliation, the model also predicts the Care/Harm foundation, which is however not present in the dataset likely due to the emphasis given by the triggers "attack" and "insult".

7 Conclusions

We presented ME^2-BERT, a novel holistic framework leveraging emotions and events for moral foundation prediction. The promising performance of ME^2-BERT supports our hypothesis of interdependence among emotions, events and moralities. Nevertheless, further investigation is needed to deeply understand the multifaceted nature of the interactions arising among these fundamental aspects, for developing more effective and interpretable LLM-based framework for moral foundation prediction and related tasks.

8 Limitations

ME^2-BERT is designed to be in principle independent from the particular PLM employed in the encoding phase as well as from the dataset used to fine-tune the PLM. Therefore, other or more advanced BERT-like models could easily be used in alternative to BERT—indeed, we also experimented with Sentence-Transformer models, as reported in Appendix F.2. Less straightforward would be the replacement of E2MoCase since, to date, there are no alternatives to it that share the same features incorporating annotations at level of emotions, events and moral foundations at the same time. Nonetheless, some choices in ME^2-BERT were influenced by the characteristics of E2MoCase. First, while our strategy for categorizing emotions has a strong theoretical foundation (Plutchik, 2001), the label's selection process was constrained by the emotion values available in E2MoCase. For other datasets, different choices might prove more effective, especially given that the exact nature of interactions between morality and emotions in psychology remains unclear and warrants further discussion. Second, E2MoCase exhibits a domain shift between event-free and event-based paragraphs, making it well-suited for our domain adaptation strategy. However, this choice may not always be optimal for the model's effectiveness and requires further investigation on other datasets. Though, as shown in our experimental section, our event-based domain identification strategy also proved effective for the MoralEvent dataset, where ME^2-BERT achieved better performance than BERT-base. Overall, the integration of events and emotions depends on the characteristics of the fine-tuning dataset and is subject to the expertise of domain specialists.

We employed a straightforward strategy for selecting emotion categories for our contrastive learning strategy, where each text is linked to a single emotion (the one with the highest score in E2MoCase). However, since emotions are fluid and texts often convey a mix of emotions (Cambria et al., 2012), a more effective strategy would involve utilizing a wider range of emotions. We aim to investigate this approach in future works.

Finally, we cannot definitively answer whether emotions and events are what we need for moral foundation prediction. We believe this matter should first be addressed using tools from moral psychology and then integrated into the development of AI tools for moral tasks. By contrast, our study offers a complementary perspective by starting with AI tools to explore whether emotions, events, and moralities are interdependent and can be effectively leveraged for moral tasks. Another factor preventing a conclusive answer is the lack of reliable datasets annotated simultaneously with morality, emotions, and events. Nevertheless, we believe our work serves as a starting point for future interdisciplinary research in this area.

9 Ethical remarks

The intended application of our framework is to show that moralities, emotions and events are interconnected aspects that can be exploited through PLMs for addressing the moral foundation prediction task. Nonetheless, biases embedded in the training data may propagate through the model we experimented with, making it essential to carefully consider its deployment to prevent the reinforcement of harmful stereotypes or biases. Ensuring ethical alignment requires ongoing evaluation.

Moreover, we acknowledge that our framework employs psychological theories (i.e., MFT for moralities and Plutchik's wheel for emotions), which however are not immune from limitations in terms of cross-cultural generalizability (Iurino and Saucier, 2020). We also hypothesize that the issues discussed in DiBerardino and Stark (2023); Stark (2023) regarding the challenges of universality and unambiguous detection of emotions might also arise in the detection of morality, and more broadly, in AI systems designed to integrate psychosocial dimensions. Nonetheless, while acknowledging that the use of MFT or any other theory may carry cultural biases and potentially conflicts with particular cultural contexts, we emphasize that the conceptual design of our proposed framework is general and not inherently dependent on the use of a particular moral theory.

Acknowledgements

A.T., resp. C.M.G, is supported by project "Future Artificial Intelligence Research (FAIR)" spoke 9 (H23C22000860006), resp. project SERICS (PE00000014), both under the MUR National Recovery and Resilience Plan (NextGenerationEU). L.Z. is supported by project PRIN2022 "AWESOME" (H53D23003550006). L.Z. and C.M.G. were visiting the University of Lausanne mostly during the development of this work.

References

Marwa Abdulhai, Gregory Serapio-García, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333.

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

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*.

Mohammad Atari, Jesse Graham, and Morteza Dehghani. 2020. Foundations of morality in iran. *Evolution and Human Behavior*, 41(5).

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79.

Kevin L Blankenship, Traci Y Craig, and Marielle G Machacek. 2021. The interplay between absolute language and moral reasoning on endorsement of moral foundations. *Frontiers in Psychology*, 12.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In *Cognitive behavioural systems: COST 2102 international training school*. Springer.

C Daryl Cameron, Kristen A Lindquist, and Kurt Gray. 2015. A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and social psychology review*, 19(4).

Stéphane Clinchant, Gabriela Csurka, and Boris Chidlovskii. 2016. A domain adaptation regularization for denoising autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

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, NAACL-HLT*.

Nathalie DiBerardino and Luke Stark. 2023. (anti)-intentional harms: The conceptual pitfalls of emotion ai in education. Association for Computing Machinery.

Matthew Feinberg and Robb Willer. 2015. From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, 41(12).

Jeremy Frimer. 2019. Moral foundations dictionary 2.0.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59).

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods*, 50.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference*. Springer.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5).

Candida M. Greco, Lorenzo Zangari, Davide Picca, and Andrea Tagarelli. 2024. E2mocase: A dataset for emotional, event and moral observations in news articles on high-impact legal cases. *Preprint*, arXiv:2409.09001.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1).

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4).

Jonathan Haidt et al. 2003. The moral emotions. *Handbook of affective sciences*, 11(2003).

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020a. Moral foundations twitter corpus: A collection of 35k tweets

annotated for moral sentiment. *Social Psychological and Personality Science*, 11(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. 2020b. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8).

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*.

Zachary Horne and Derek Powell. 2016. How large is the role of emotion in judgments of moral dilemmas? *PloS one*, 11(7).

K. Iurino and G. Saucier. 2020. Testing measurement invariance of the moral foundations questionnaire across 27 countries. *Assessment*, 27.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

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

Enrico Liscio, Alin E Dondera, Andrei Geadau, Catholijn M Jonker, and Pradeep K Murukannaiah. 2022. Cross-domain classification of moral values. In *2022 Findings of the Association for Computational Linguistics: NAACL 2022*.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Alejo Lopez-Avila and Víctor Suárez-Paniagua. 2023. Combining denoising autoencoders with contrastive learning to fine-tune transformer models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Negar Mokhberian, Frederic R. Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE.

Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. 2024. Measuring moral dimensions in social media with mformer. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM*. AAAI Press.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Association for Computational Linguistics.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4).

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: A fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. Association for Computing Machinery.

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*.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25. Academic Press.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1).

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11.

Luke Stark. 2023. Artificial intelligence and the conjectural sciences. *BJHS Themes*, 8.

Melike Tekin and Sönmez Ekici. 2023. The relationship between emotions and morality. *Frontiers in Psychology*.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Giuseppe Ugazio, Claus Lamm, and Tania Singer. 2012. The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12(3).

Ibo Van de Poel and Lambèr Royakkers. 2023. *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*. Association for Computing Machinery.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7.

Xinliang Frederick Zhang, Winston Wu, Nicholas Beauchamp, and Lu Wang. 2024. MOKA: Moral knowledge augmentation for moral event extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Notations

Frequently used symbols and descriptions used throughout the main paper are summarized in Table 5.

B Events encoding

The events in E2MoCase are provided as JSON objects, representing each event as a tuple of its corresponding trigger words and involved entities. To provide BERT with contextual information and to enable it to extract meaningful text embeddings, we convert the JSON data into descriptive texts that detail the trigger words and involved entities of each event, as illustrated in Fig. 9.

C Plutchik's Wheel of Emotions

Plutchik's Wheel of Emotions (Plutchik, 2001) is a model that categorizes human emotions into eight primary types: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*. These emotions are arranged on a wheel to represent their relationships and intensities, with similar emotions placed close to each other and opposites across from one another. The wheel, illustrated in Fig. 10, also shows how primary emotions can blend to form complex emotions. This model provides a systematic approach to understanding emotional responses, supporting various fields, including psychology, artificial intelligence, and natural language processing.

D Data description

Following (Guo et al., 2023; Preniqi et al., 2024), we preprocess all datasets by removing URLs, hashtags, and non-ASCII characters, replacing user mentions with "@user", and converting emojis to their textual equivalents. Next, we describe the data used for fine-tuning and evaluation.

D.1 Fine-tuning data

In our main experiments, we fine-tuned ME^2-BERT on the E2MoCase dataset, specifically its E2MoCase_full version, which contains text segments with and without events. Furthermore, we also used the MoralEvent dataset to validate our event-based domain identification strategy.

E2MoCase. E2MoCase_full (Greco et al., 2024) is designed to include a diverse range of news media reports on popular cases involving religious, political, gender, racial, and media biases. E2MoCase is automatically annotated at paragraph level with emotional tone, moral traits, and events. Emotional and moral traits are provided with a strength score ranging from 0 to 1. Events are presented as JSON objects, as shown in Fig. 9, containing the trigger word, involved entities, and their roles within the text context. Note that a single paragraph can encompass zero, or more emotions, moral traits, and events. The dataset consists of 97,251 paragraphs, of which 50,975 contain events, while the remaining paragraphs are without events. The average number of token is 210 (Word Piece tokenizer (Devlin et al., 2019)). The average alternace rate of paragraphs with and without events within a news article is about 66%, indicating a significant presence of event-free paragraphs that complement the eventful ones. Including paragraphs without identified events provide essential background information, explanations, or transitions

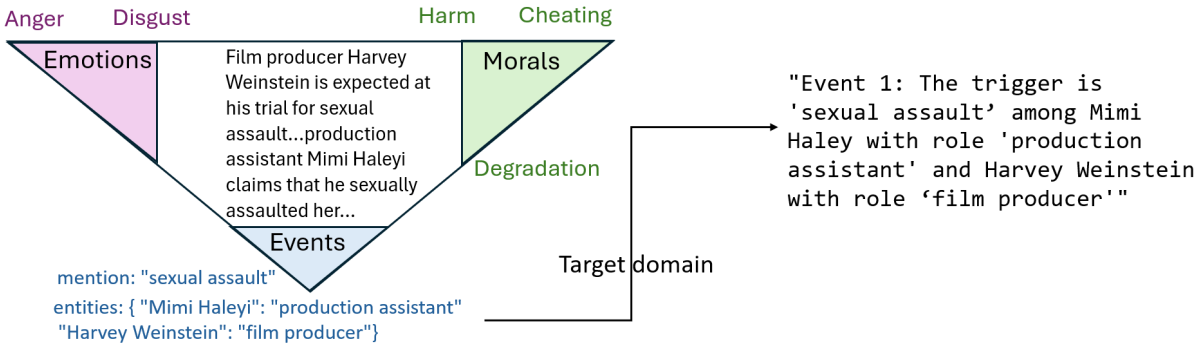


Figure 9: Event encoding in descriptive text.

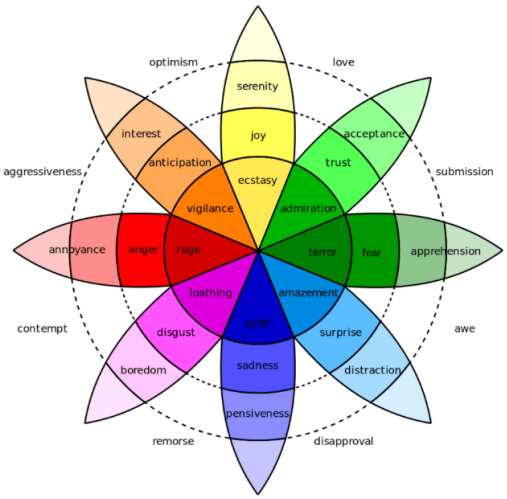


Figure 10: Plutchik's Wheel of Emotions

between eventful content, enabling us to exploit a bigger corpus of data. Furthermore, since emotions and moralities in E2MoCase are associated with strength scores, we applied a threshold of 0.5 to determine the presence of an emotion or morality.

MoralEvent. The MoralEvent dataset (Zhang et al., 2024) contains structured event annotations from 474 news articles by diverse US media outlets. It consists of 5,494 event annotations and is unique in that annotations are conducted on multiple news articles about the same story, allowing for analysis of differences in how news outlets of different ideologies report moral events. The dataset also includes moral annotations, capturing implicit participants in moral actions. The annotations cover a wide range of news sources and diverse entity types, including People, Organizations, Geo-Political, and Others. The dataset provides detailed information on the types of reported events and the moral foundations reflected in the interactions among participating entities.

D.2 Evaluation data

Moral Foundation Twitter Corpus (MFTC). The MFTC (Hoover et al., 2020b) consists of 35,108 tweets from seven discourse domains, chosen based on their moral relevance and popularity among Twitter users. The domains include All Lives Matter (ALM), Black Lives Matter (BLM), Baltimore protests, the 2016 Presidential election, hate speech and offensive language, Hurricane Sandy, and the #MeToo movement. Each tweet in the Moral Foundations Twitter Corpus (MFTC) was hand-annotated with the five MFT categories, i.e. Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Purity/Degradation.

Moral Foundation Reddit Corpus (MFRC). The MFRC (Trager et al., 2022) comprises 16,123 English Reddit comments from 12 subreddits, chosen based on expected moral content, activity level, and diversity. The subreddits are organized into three categories: US politics, French politics, and everyday moral life. It is hand-annotated for eight MFT foundations, namely: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, Liberty/Oppression, Equality/Inequality and Proportionality/Disproportionately, Thin Morality and No Moral. We consider the first five foundations, and combine proportionality and equality under the fairness foundation.

The Extended Moral Foundation Dictionary (eMFD). The eMFD (Hopp et al., 2021) contains 35985 news articles on a variety of topics, which was manually annotated by 557 annotators. The annotation involved text spans and their embodied moralities according to the MFT principles.

Note that we refer to the dataset as eMFD; other studies (Mokhberian et al., 2022; Nguyen et al.,

2024) refer to this dataset as MNFC (Moral Foundations News Corpus). Here, we follow Guo et al. (2023) in referring to it as eMFD, but we clarify that we are referring to the dataset and not the dictionary constructed by Hopp et al. (2021).

E Details on experimental methodology

E.1 Implementation details

We implemented ME^2-BERT, BERT-E2MoCase, BERT-MFRC, BERT-MFTC and BERT-MFRC using PyTorch library.[2] For DAMF,[3], MoralStrength[4] and DDR,[5] we used their publicly available source code. For MoralStrength, we employ the best performing model (on average) according to Araque et al. (2020). For MoralBERT,[6] Llama-3.1,[7] Gemma-2[8] and Mistral-Nemo,[9] we use the models available on HuggingFace. [10]

E.2 Prompt construction for LLMs

The same prompt is constructed for all models in a zero-shot setting, providing the paragraph and the list of possible moral traits as options. Specifically, the following prompt is provided:

Prompt

Given the following news article paragraph:

{paragraph}

Categorize the text's moral traits as 'neutral' or according to the moral foundation theory with the following: *{list of all possible morals traits}*.
Your response must follow the following pattern: [<list of detected moral traits>]. Your response must contain just the list of the detected moral traits, do not add any additional word or introductions.

E.3 Hyper-parameters

Fine-tuning for all models was performed over 10 epochs with a batch size of 8. For all BERT-based models, we set a maximum sequence length of 256 due to the length of paragraphs in E2MoCase (about 210 tokens per paragraph on average, with high variability), with padding to the maximum length. We adopted the token [CLS] as text embeddings. For ME^2-BERT, we applied noise to the auto-encoder by using the dropout technique (i.e., by randomly setting to zero some values with probability p). We performed a grid search to select the best learning rate and dropout values, i.e., 5.0E-5 and 0.3, respectively. The margin m of the contrastive function dynamically varied from 1 to 0 using an exponential function. The term λ_{dom} was also adjusted exponentially from 0 to 1 (Ganin et al., 2016; Guo et al., 2023). BERT-E2MoCase, BERT-MFRC, BERT-MFTC, and BERT-MFTRC were fine-tuned using the same hyper-parameter setting as ME^2-BERT. All other competing methods were carried out using the default hyperparameters specified in their corresponding source code. Considering LLMs, each model is asked to respond to the prompt described in Sect. E.2 with the role of *expert in Moral Foundation Theory*. We set the models' *temperature* to 1.0E-2, so as to reduce randomness and ensure the selection of the most probable tokens for more precise and deterministic responses. We ran each model five times and reported its average performances.

E.4 UMAP

The Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) graphs were plotted using the following hyper-parameters: number of neighbors and minimum distance equals to 80 and 0.5, resp., and using cosine similarity as a distance measure. The plots in Figs. 3 and 7 were generated using 10,000 random samples.

E.5 Environment

We conducted all the experiments on a Linux machine (OS Ubuntu 20.04.06 LTS), equipped with 256GB of memory, processor Intel(R) Xeon(R) Gold 6248R CPU, 3.00GHz and GPU NVIDIA A30 with 24GB memory and CUDA version 11.8.

F Additional Results

F.1 Results on the eMFD dataset

We report additional results achieved by ME^2-BERT on the eMFD dataset. Figure 11 shows the F1 scores yielded by ME^2-BERT when trained in multi-label and single-label setting (ME^2-BERT and ME^2-BERT-SL, respectively), against the best competing method (DAMF and DAMF-SL). Similary to the performance observed in Fig. 5 on the MFRC and MFTC datasets,

[2] https://pytorch.org/
[3] https://github.com/fionasguo/DAMF/tree/master
[4] https://pypi.org/project/moralstrength/
[5] https://github.com/USC-CSSL/DDR
[6] https://github.com/vjosapreniqi/MoralBERT
[7] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
[8] https://huggingface.co/google/gemma-2-9b-it
[9] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[10] https://huggingface.co/

the multi-label models tend to outperform their single-label counterparts also in this case.

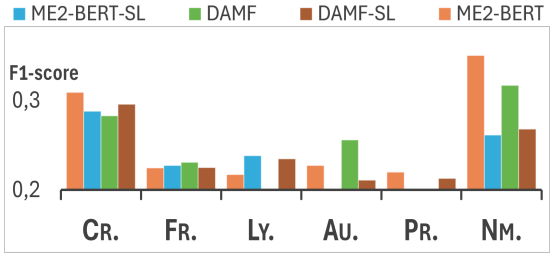


Figure 11: Single-label classification on the eMFD dataset.

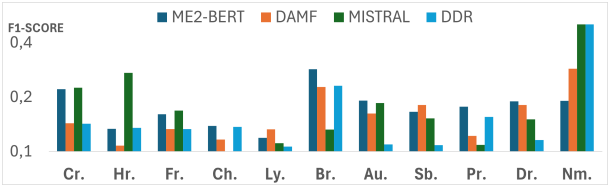


Figure 12: Detection of polarity of moral dimenisions on the eMFD dataset.

Figure 12 shows the results of ME^2-BERT when compared with the best competing methods of each category (LLMs, lexicon and BERT-based) on the eMFD dataset. In this case, ME^2-BERT is the second-best method with an average F1 score of 0.170, outperformed only by Mistral-Nemo, achieving 0.172 F1-score. Overall, Mistral-Nemo appears to be highly effective at predicting the Harm and non-moral polarities. ME^2-BERT, on the other hand, is highly effective in predicting the moralities Betrayal (Br.), Purity (Pr.), and Degradation (Dr.). Note that the MFRC dataset does not contain information on the polarity of moral foundations.

F.2 Fine-tuning with Sentence-Transformers

In our main experiments, we used *bert-base-uncased* to ensure fairness with PLM-based competitors, which refer to the aforementioned version of BERT. However, as highlighted in Sect. 8, our framework is PLM-agnostic. Table 4 shows the results of our framework when fine-tuned with different Sentence-Transformers (Reimers and Gurevych, 2019) as backbones, such as *bert-base-nli-mean-tokens* (ME^2-SBERT) *sentence-t5-large* (ME^2-T5), *nli-roberta-base-v2* (ME^2-RoBERTa), and *all-mpnet-base-v2* (ME^2-MPNET). Sentence-Transformers can capture semantically meaningful embedding, making them suitable architectures for sentence similarity tasks. Note that the Sentence-Transformer models were run using 128 as the maximum sequence length for efficiency purposes, since some models, like T5, have a large number of parameters (355M). ME^2-BERT was fine-tuned using 256 as maximum sequence length.

	Cr.	Fr.	Ly.	Au.	Pr.	Nm.	AVG
MFRC							
ME^2-BERT	0.636	0.585	0.345	0.490	0.363	0.669	0.515
ME^2-SBERT	0.629	0.583	**0.359**	0.483	0.352	0.660	0.511
ME^2-T5	**0.661**	**0.604**	0.345	**0.503**	**0.384**	**0.715**	**0.535**
ME^2-RoBERTa	0.630	0.575	0.289	0.472	0.346	0.685	0.500
ME^2-MPNET	0.596	0.558	0.270	0.427	0.295	0.563	0.452
MFTC							
ME^2-BERT	0.688	0.673	**0.551**	0.521	**0.453**	0.546	0.572
ME^2-SBERT	0.692	0.669	0.548	0.515	0.451	0.516	0.565
ME^2-T5	**0.710**	**0.678**	0.549	**0.564**	0.433	**0.585**	**0.586**
ME^2-RoBERTa	0.681	0.642	0.535	0.506	0.438	0.573	0.562
ME^2-MPNET	0.673	0.616	0.531	0.467	0.429	0.528	0.541
eMFD							
ME^2-BERT	0.309	0.225	0.217	0.227	0.220	0.373	0.262
ME^2-SBERT	0.299	0.238	0.218	0.242	**0.243**	0.288	0.255
ME^2-T5	0.292	**0.272**	0.247	0.251	**0.243**	0.212	0.253
ME^2-RoBERTa	**0.324**	0.265	**0.254**	0.265	0.241	**0.423**	**0.295**
ME^2-MPNET	0.323	0.269	**0.254**	**0.271**	0.237	0.397	0.292

Table 4: F1-scores of ME^2-T5, ME^2-RoBERTa, and ME^2-MPNETmodels fine-tuned on MFRC, MFTC, and eMFD datasets. Best results for each moral foundation in bold, second-best underlined.

Overall, the best performing method is ME^2-T5 with average F1-score of 0.458 across all datasets, followed by ME^2-RoBERTa (0.450) and ME^2-BERT (0.450). ME^2-T5 shows outstanding performance on social datasets like MFRC and MFTC, but its performance decreases on eMFD. By contrast, ME^2-RoBERTa and ME^2-MPNET perform poorly on MFRC and MFTC but achieve the highest scores on eMFD. This highlights the strong differences between these datasets, as evidenced by their feature distributions in Fig. 3b. Other models, such as ME^2-BERT and ME^2-SBERT, exhibit more balanced performance across all datasets. Notably, ME^2-BERT achieves comparable, and in some cases even better, performance than ME^2-SBERT. We believe that the good performance achieved by ME^2-BERT w.r.t. SBERT are due to the task-specific fine-tuning process, as the relative performance of a model can also depend on the similarity of the pretraining and target tasks (Peters et al., 2019).

Notation	Description
$\mathcal{D}$	Generic dataset with pair of samples $(d_i, \mathbf{y}_i)$.
$\mathbf{y}_i$	Real-value scores associated to MFT dimensions.
d_i	Generic text segment.
$\mathcal{T}$	Space of all possible input texts.
$\mathcal{D}_s$	Labeled source domain dataset.
$\mathcal{D}_t$	Unlabeled target domain dataset.
$d_i^{(s)}, d_j^{(t)}$	Text segment from the source and target domains, respectively.
n	Number of text segments.
n_s, n_t	Number of text segments of source and target domains, respectively.
$y^{(s)}, y^{(t)}$	Ground-truth moral foundation scores for source and target domains, respectively.
$\mathbf{y}_i$	Real-value vector for MFT dimensions.
f	Moral foundation classification function.
g	Domain classification function.
σ	Sigmoid activation function.
$\mathbb{P}, \mathbb{Q}$	Marginal and Conditional distributions, respectively.
ψ	Noise generated with Bernoulli distribution.
p	Probability of dropout.
$\odot$	Element-wise multiplication.
$\mathbf{x}_i, \mathbf{x}'_i$	BERT embedding of the i-th text segment, and its corrupted version, respectively.
$\mathbf{x}_i^{(oob)}$	Out-of-the-box BERT embedding
$\hat{\mathbf{x}}_i$	Reconstructed BERT embedding.
$\mathbf{h}_i$	Bottleneck representation learned by the encoder.
f_{enc}, f_{dec}	Encoder and decoder functions of E-DAE.
$\mathbf{h}_p, \mathbf{h}_r$	Positive and negative samples for the contrastive learning function.
λ_{dom}	Weighting parameter for adversarial loss.
$\mathcal{L}_{\text{MSE}}$	Mean Squared Error (MSE) loss for reconstruction.
$\mathcal{L}_{\text{tr}}^{(s)}, \mathcal{L}_{\text{tr}}^{(t)}$	Souce and Target triplet margin losses for contrastive learning, respectively.
$\mathcal{L}_{\text{con}}$	Contrastive loss.
$\mathcal{L}_{MF}$	Cross-entropy loss for moral foundation prediction.
$\mathcal{L}_{ADV}$	Loss for adversarial domain classification.
$\mathcal{L}_{E-DAE}$	Loss of the E-DAE module.
$\mathcal{L}$	Overall Loss.
$\tilde{\mathbf{x}}_i$	Gated representation.
$\mathbf{f}_i, \mathbf{q}_i, \mathbf{o}_i$	Forget, input, and output gates.
$y_i^{(D)}$	Ground-truth label indicating which domain the input segment d_i belongs to.
T	Maximum sequence length of the model.
n_b	Number of samples in a batch.
d_b	Hidden dimension of each neural model outside the BERT encoder.
d_B	Hidden dimension of BERT.
BERT-E2MoCase	BERT model fine-tuned on E2MoCase dataset.
BERT-MFTC	BERT model fine-tuned on MFTC dataset.
BERT-MFRC	BERT model fine-tuned on *MFRC* dataset.
BERT-MFTRC	BERT model fine-tuned on the aggregation of the MFRC and MFTC datasets.
ME^2-BERT w/o A.	ME^2-BERT without the Adversarial Learning module.
ME^2-BERT w/o A-C.	ME^2-BERT without the Contrastive and Adversarial Learning modules.
ME^2-BERT w/o C.	ME^2-BERT without the Contrastive Learning module.
ME^2-BERT w/o G.	ME^2-BERT without the Gate machanism.

Table 5: Table of notations and model names