Research Report **meinung**

What makes us human? One answer is ethics and morality. The TUMJA project meinung analyses moral stances in text. It builds upon an existing mapping of words to their moral meanings, and combines this with the large language model BERT. As such, meinung improves existing methods of measuring morality by extending the scope of applicable words, i.e. moral vectors can now be calculated for almost any word – which also improves performance for short texts. Additionally, the approach of meinung allows differentiating identical words in different contexts and thus different meanings.

Team	Viktoria Obermeier Yunqing Wang Maximilian Frank Tim Knothe Julius Mankau Nikola Martinov Staykov Christian Nix
Tutors	Eva-Madeleine Schmidt Stefan Engels
Supervisor	Prof. Dr. Claudia Klüppelberg Prof. Dr. Martin Werner

Preface by the Supervisors

Prof. Dr. Claudia Klüppelberg und Prof. Dr. Martin Werner

Information provided by newspaper texts expresses and influences moral perceptions of our society. Our team "meinung" had the ambition to understand possible changes in moral perceptions over the years and chose environmental news as one prominent example. When natural disasters like floods, heatwaves, earthquakes, and hurricanes strike, the media report immediately. Moreover, questions are raised about responsibility and blame. Could the disaster have been prevented, was the response of government and disasters services adequate? This way, often, a natural disaster leads to moral questions and to discussions about climate change, including accusations of neglect of sustainability issues by politicians.

The team members drew on different scientific methods to study word connotations by building a metric based on counts of negative and positive words. Acknowledging the complexity of the problem of assessing moral perceptions within society, our team used five different base criteria: authority, care, fairness, loyalty and sanctity. Their study also incorporated considerations of context awareness as well as non-explicit context. The fact that language changes naturally over time is also taken into account. An algorithm has been designed to implement these issues for analyzing moral connotations in texts and compared favorably to the extended Moral Foundations Dictionary.

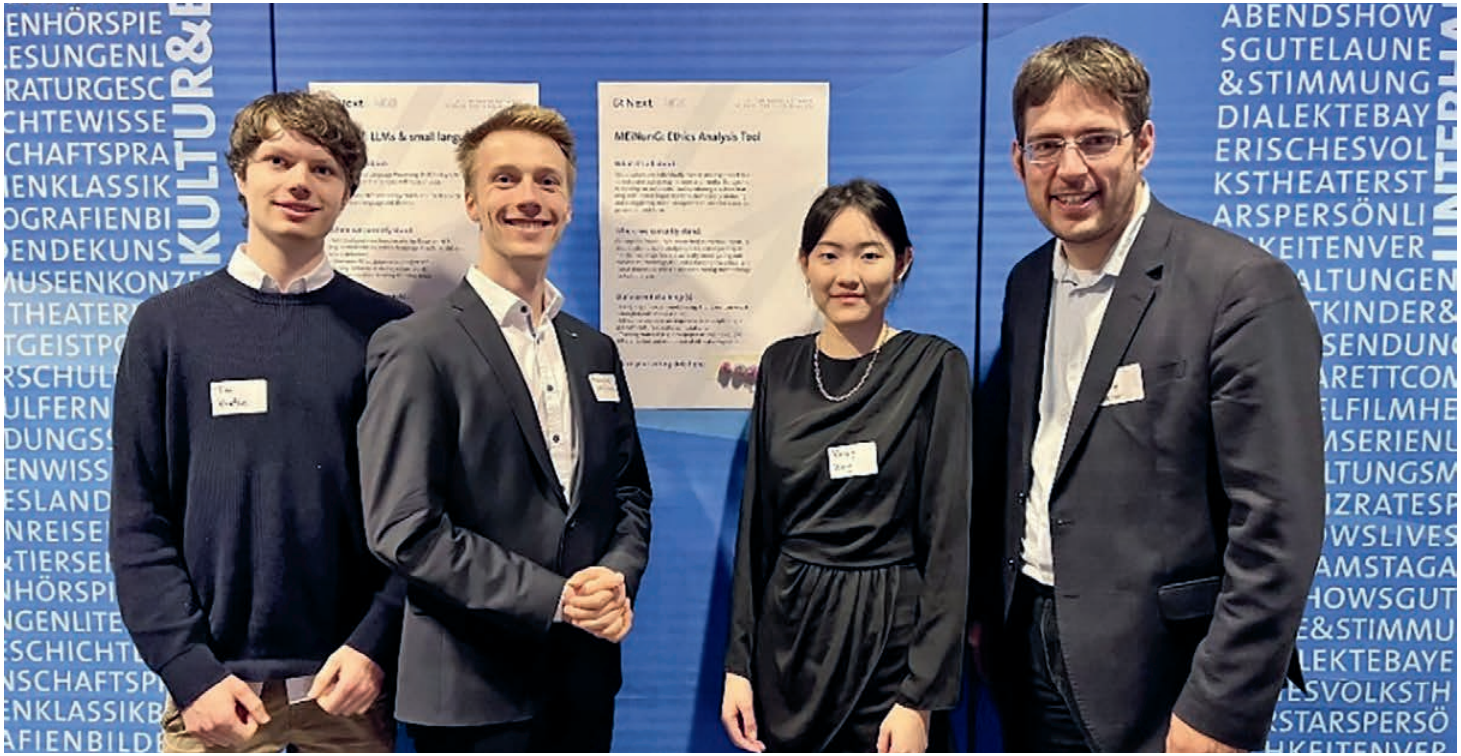
The team succeeded in a demanding project of interdisciplinary research at the intersection of sociological, philological and computer science contributing to the discussion of the ethical foundations of our democracy.

Beyond the technical achievements, what stood out most was the spirit of cooperation and the shared vision within the team. Despite different backgrounds and skills, they were united by a common goal, which led to valuable project results.

As supervisors, we had the privilege of observing a team of young students developing from highly motivated singletons in different study programs into a joint research team, whose interdisciplinary members brought in their diverse personalities and disciplinary strengths to successfully conclude a unique and socially important project within the given deadline. Each member brought unique perspectives and expertise to address the challenges of complex text- searches in answering questions about how our society changes moral perceptions.

One of the best TUMJA moments was the invitation of team members to the BR Forum "AI for Media: Science meets Journalism" organized by Bavarian Public Broadcasting Service (see photo). In this meetup, they presented their initial results and got into discussion with practitioners from the press.

As supervisors of the team, we are proud of you and all your achievements. Working with you has always been a pleasure.


ABENDSHOW
SGUTELAUNE
&STIMMUNG
DIALEKTEBAY
ERISCHESVOL
KSTHEATERST
ARSPERSÖNLI
ALTUNGEN
SENDUNG
ELFILMHE
LTUNGSM
AMSTAGA
HOWSGUT
E&STIMMU
LEKTEBAYE
ESVOLKSTH
STARSPERSÖ
KULTUR&
MEANinGful Ethics Analysis Tool

Can We Measure Morality?

The TUMJA research project *meinung* wants to enhance the current methods of morally analyzing text. But why is moral analysis a very relevant topic – if not crucial for democracy? Can we really push moral analysis beyond the theoretical analysis of dry text? Yes, we can. Because text is not just a form of expressing thoughts. Especially with newspapers, text actually becomes a mirror of society.

Floods, heatwaves, earthquakes, and hurricanes – natural disasters seem to be occurring with increasing frequency and severity. When disasters like these strike, the media responds swiftly, often becoming the lens through which the public experiences the chaos and aftermath. Headlines flash across news channels, websites, and social media, capturing not only the scale of the destruction but also the stories of those affected. While images of flooded streets, submerged homes, and displaced families dominate the front pages, there's often another narrative at play – one that extends beyond the immediate devastation. This narrative is about morality, responsibility, and blame: Who is at fault for the disaster's impact? Could it have been prevented? And what do these events reveal about our societies, our governments, and the way we treat our planet?

The ways these questions are framed in the media can shape public perception dramatically. In the United States, for instance, the response to Katrina was as much about the failure of federal relief efforts as it was about the hurricane itself. Even questions of race, class, and inequality surfaced, turning a natural disaster into a moral and political reckoning. Meanwhile, in Europe, coverage of extreme weather events often shifts quickly to debates over climate change, sustainability, and the collective responsibility of humanity for the increasingly volatile environment.

As we can see, news media bear a great responsibility in society, forming opinions and ultimately having an effect on political decisions. One thing which mainly influences how the news is perceived in the media is which words have been used – and we need to consider not only the words themselves, but also the contexts into which they are placed as well as their target audience. Unfortunately, many newspapers create captivating headlines to attract a large audience – at the cost of neutrality. So, it is essential to measure how words have been used in a newspaper article to increase transparency, which is quite important for a working democracy.

Word Connotations: There is a big difference in using words and phrases with similar meanings but different connotations. Speaking of environmental topics, you could describe agricultural farming as "a cultivation of nature to maximize the growing of food" – or, by contrast, "the occupation of natural space to satisfy human greed." When analyzing such varying connotations, and whether a news article expresses a positive or a negative viewpoint, one could easily assign each word a number on a scale from positive (e.g., 1) to negative (e.g., -1), and sum the values up. For example, in the positive example above, we have "cultivation" and "maximize" – two positive words, i.e., score +2. The negative example contains "occupation" and "greed" – two negative words,

i.e., score -2. However, a balanced sentence or article would yield a number near zero, like “Agriculture is a cultivation of nature to maximize the growing of food. However, in the context of rainforests it may seem less negative to say it is rather an occupation of natural space to satisfy human greed” – yielding a score of 0. By using this metric now, one could not only classify news articles into positive or negative, but also observe the newspapers’ orientation over time. In fact, there already exists such a score, called the extended Moral Foundation Dictionary. As measuring morality and word connotations is quite complex, it does not have just one dimension as in the examples above (i.e., negative to positive), but actually five different dimensions: authority, care, fairness, loyalty, and sanctity.

Context Awareness: One downside of the above-described method is that it is unaware of any context. For example, the expression “love”, as related to parents, has quite a different meaning than “love” as related to ice cream – even though it is the same word. That’s where natural language processing comes into play: It translates all words into a mathematical construct called vectors – you can think of them as virtual words. So, in very simplified form, a sentence like “I love ice cream” is turned into “I love-as-in-like ice cream“ and “I love my parents” is turned into “I love-as-in-relations my parents.” Now, you can compare the words again, because “love-as-in-like” is a different word than “love-as-in-relations.” In fact, natural language processing doesn’t compare words like we humans do, but rather using the aforementioned vectors and numbers.

Non-explicit Context: However, context alone is not sufficient either. Even expressions “I like trees” or “I love my parents” have very different meanings across cultures. For example, in Eastern countries, parents enjoy a much greater authority beyond a child’s legal age than in Western countries. People in the mid-latitudes have a very different understanding of the robustness of trees than in African countries. For example, a robust tree in Germany is associated with an oak, whereas in Africa it is mainly associated with robustness against drought.

Time-Factor: Even if you account for all aforementioned factors during the analysis of the connotations in news articles – language still changes over time. The German word “geil” (colloquial for awesome) had a very different meaning a few decades ago: horny. And language now seems to change faster than before. For analyzing moral connotations in texts, it is thus almost essential to find an automated algorithm, so that the analysis can be conducted faster, more efficiently, and – most importantly – neutrally.

This is in essence what the TUMJA research project *meinung* is about: Enhancing methods of moral analysis in text. In this, we aim to contribute to existing research at the intersection of sociological and philological research as well as computer science. Moreover, measuring morality facilitates the assessment of opinionating articles and thus serves as a factor of transparency – ultimately an important prerequisite for democracy. ■

Research Report – meinung

Enhancing the extended Moral Foundation Dictionary with BERT

Table of contents:

Introduction

Moral values, as a product of human civilization and communal living, continually evolve and adapt along with changes in the spirit of the times. Most research examining moral content in text has been based on the practical application of moral foundations theory [1]. This suggests that individuals across different cultures and societies share five innate and universal moral foundations, each with their positive/negative poles as virtue and vice: care/harm (incl. sympathy, compassion, and nurturance), fairness/cheating, loyalty/betrayal, authority/subversion (involves concerns about traditions and maintaining social order), and sanctity/degradation (involves moral disgust and spiritual concerns related to the body). Using moral foundations theory as a framework, dictionary-based approaches have been developed to analyze moral content. The methods used therein focus on identifying the frequency at which keywords related to moral foundations appear in a text [2].

History of Moral Measuring

Graham et al. [3] created the first dictionary based on moral foundations theory. This dictionary was constructed through the manual selection of words from thesauruses and conversations with colleagues, which were chosen to represent the upholding or violation of specific moral foundations. After some research, the Moral Foundations Dictionary (MFD) [4] was first implemented to study differences in moral language, for example in religious texts [3]. In essence, the MFD maps words with moral values to one or more moral foundations, but most often only one.

While the MFD provides a straightforward, word-count-based method for extracting moral content from text, several concerns [5–7] have been raised regarding its theoretical validity, practical utility, and scope. Hopp et al. [8] summarized these concerns into three categories:

Validity and Generalizability. The MFD is constructed using lists of moral words, which were deliberately selected by a small group of experts [3]. This approach raises concerns about the dictionary's ability to accurately capture intuitive moral processes in the general population, and thus invites criticism of its validity.

Categorization Limitations. The MFD and similar tools rely on a binary approach, where each word is assigned fully to a moral foundation – but this clearly does not allow for any scalar differentiation (e.g. slightly fair). This rigid classification constrains the dictionary's ability to reflect the natural variation in moral information and its contextual meanings across diverse situations.

Simplified Representation. The methods used for the MFD conceptualize text as so-called "bags of words" [9], which significantly limits their ability to capture the relational structure of moral acts, such as identifying the actors involved, the nature of the actions, and the underlying reasons for their occurrence.

To address these limitations, Hopp et al. [8] developed the extended moral foundations dictionary (eMFD). In contrast to previous approaches, the creation of the eMFD involved a web-based, hybrid content annotation platform, known as the Moral Narrative Analyzer (MoNA). Annotators were instructed to identify five predefined moral foundations within news articles, which were sourced from major media outlets. Prior to annotation, the texts were preprocessed through tokenization, stop-word removal, and part-of-speech tagging, which is a form of tagging words with their grammatical function (e.g. nouns, verbs, …).

A total of 2995 articles were annotated, and words or phrases were assigned probabilities for each moral foundation based on annotation frequency. So, words were not only assigned to one or more moral foundations, but also had different degrees of intensity in each moral foundation. This intensity is measured on continuous scales between -1 and +1, e.g. +1 for care, -1 for harm, 0.6 for rather care, and so on. To ensure reliability, lexical items were filtered and retained only if they occurred multiple times across different annotators and contexts. Additionally, sentiment analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner) was employed to classify words into "virtue" or "vice" categories, thereby capturing their moral valence.

Besides creating the dictionary, Hopp et al. also validated it on a newspaper articles dataset. The accompanied Python library eMFDScore enables moral analysis through bag-of-words and syntactic dependency parsing methods. The eMFD was validated through statistical tests, comparing it to previous dictionaries (like MFD 2.0). With a total of 3270 words, the eMFD provides a robust tool for analyzing moral connotations in text and serves as a base for further research.

Research Questions

One flaw with static word lists is that any change of meaning due to different contexts is normally not reflected in the measures for each moral foundation. For example, the word "love" in "I love ice cream" has a different meaning than in "I love my parents." Despite the eMFD's attempt to include contextuality in calculating moral connotations of words, the moral values are combined into a single moral vector for each word. This lack of differentiation also applies to the eMFD. Thus, we investigate how the eMFD can be improved to account for different meanings of the same words in different contexts (RQ1).

Second, as the moral foundation dictionary was created statically, and language is constantly changing, a static version of the eMFD might not reflect all words with a moral connotation. Thus, we investigate how the eMFD can be extended to measure moral connotations of words which are currently not within the eMFD (RQ2). This enables moral measurement of words beyond its limits. Summarizing the above, our research questions are as follows:

- *RQ1:* How can the eMFD be improved to also account for different meanings of words resulting from different contexts?
- *RQ2:* How can the eMFD be extended to measure moral connotations of words currently not within the eMFD?

To answer these questions, we combine the eMFD with a large-language model, which models a superset of the words in the eMFD. In this way we can generalize moral dimensions to other words which are not mapped to a moral vector yet.

Our moral language model allows to capture differences in semantic meanings of words by taking their context into account. This is crucial for accurate moral and semantic analysis. Additionally, our

approach is not based on a static dictionary but effectively allows for modularly extending the eMFD based on the used large-language model. And lastly, we validated our model by using difference-in-difference comparisons with the eMFD. Validation was conducted with regard to alignment with the eMFD, generalizability to words not contained in the eMFD and whether our model performs better with short texts, like those that prevail on social media. As such, we contribute an important milestone to context-aware moral measuring, which is relevant in various fields of science such as fake-news detection or neutrality analysis of text.

Related Work

There is a range of existing research related to our work. In the following, we provide a brief overview of such research.

Applications

Sagi et al. [10] performed a moral analysis on tweets connected to the U.S. government shutdown in 2013. Their main focus were the differences in moral stances between intra-community and inter-community retweets on that topic. The research also took into account how much users interacted with the actual content – besides the moral logic. Interestingly, there was a significant effect: Content was emphasized more than moral rhetoric.

Similarly, Roy et al. [11] analyzed moral sentiment in U.S. politicians' tweets about two controversial topics. The research shows a significant difference between political parties.

From both research projects, we can see that differences in moral stances can indeed be measured between groups. This is a prerequisite for verifying our research.

MFD Extensions

Rezapour et al. [12] extended the MFD [4] with a manual process involving humans. Besides that, they also applied their enhanced MFD on natural language processing tasks to test its usefulness for measuring social effects. While validating the enhanced MFD with language models, they were not used in the extension process itself. So, in contrast to our research, they do not account for different meanings of the same words in different contexts.

Due to language limitations, Cao et al. [13] created their own moral foundation dictionary by adapting moral foundation theory to the Chinese language domain. Besides manual semantic annotation, they complemented their approach with large-language models to extend their dictionary. However, their results are limited to Chinese comments on Weibo. We expect different results for English large-language models which include more versatile input sources than short-text comments.

Egorov et al. [14] proposed an orthogonal dimension to the moral foundation theory. They extended the existing five foundations each with four different sensitivities: victim sensitivity, observer sensitivity, beneficiary sensitivity, and perpetrator sensitivity. These should give a better differentiation for each moral foundation, because the perception of morality depends on each perspective.

Sagi et al. [7] use a keyword-based approach to apply the eMFD to any kind of text of a specific topic domain. This enables filtering out moral values of different topics contained in the same text. Before that, this had added additional noise to moral measurements. However, they still use the static dictionary approach of the eMFD – so while their approach discards off-topic statements, they still do not include contextuality into their measurements.

More akin to our research, Nguyen et al. [15] fine-tuned a large-language model for analyzing moral stances in newspaper articles. This model, called Mformer, was found to outperform existing approaches of moral measurement. While their language model is applicable to several text domains, they estimate moral connotations directly via the model – instead of focusing on extending the eMFD. Thus, we also use different approaches for validation than the ones used by Nguyen et al.

Unfortunately, many approaches involve manual human annotations. This doesn't scale to analyzing large text corpora. Another limitation of human-annotated moral models is noise introduced by the annotations [16]. We therefore focus on an approach which does not involve human annotation.

Methods

One of the most widely used methods to evaluate texts using the eMFD utilizes a method commonly known as bag-of-words (BoW) [8]. The BoW method counts the occurrence of words, which also have a moral vector in the eMFD. The remaining frequencies are then combined with the words' scores in the eMFD [8]. One limita-

tion of this approach is that each word in the dictionary contributes fixed moral values, regardless of its contextual usage. This context-insensitivity can lead to misinterpretation, as already noted with “love” as with ice cream or parents. We propose an approach to overcome this limitation and include context when analyzing words regarding their moral connotation.

Overview

To add contextual understanding to our moral model, we decided to combine the eMFD with a context-aware large-language model. For this, the context-aware BERT model (Bidirectional Encoder Representations from Transformers) [17] was a promising choice. BERT is a pre-trained transformer model for natural language and has found wide adoption due to its ease of use [18]. Specifically, it is aimed at generating bidirectional context-aware vector representations (i.e. embeddings) for text tokens (i.e. word fragments) [18]. So each vector of a word (i.e. embedding) uniquely codifies a certain meaning of this word. Other word vectorization models, like word2vec [19] or GloVe [20], are context-free and thus lack an essential property for our use.

First, we use the encoder of BERT to translate a sequence of tokens (i.e. word fragments) to a latent space, i.e. a high-dimensional vector space, which encodes the meaning of tokens. When using all words in a text for moral analysis, there may be a lot of noise. To isolate meaningful context-sensitive, moral directions and thus filter out noise, we reduce the amount of dimensions of the BERT latent space. This is done by creating a linear projection of the latent embedding space using singular value decomposition (SVD) [21]. During inference, token embeddings are projected into this reduced vector space to enhance model performance, following the approach of Brunton and Katz [22]. The number of dimensions of the projection is thresholded based on the method described by Donoho and Gavish [23]. We fixed the threshold at 100 for subsequent evaluations, which is the default value for SVD. Despite our resolution being quite low (reducing 768 dimensions from BERT [17] to 100 dimensions), this proved to be sufficient for our task – especially because we were only interested in five moral dimensions.

We then apply multiple linear regressions [24] to map the reduced embedding vectors to moral vectors for each word and thus decompose the context-free moral vectors from the eMFD into each of one or more context-aware versions. The regressions are modelled from the existing moral vectors in the eMFD. After calculating the moral vectors of individual words, the overall moral profile of a text is computed by averaging the individual word-level scores. For comparisons using the bag-of-words (BoW) method, the word-level scores were averaged, weighted by word frequency. Using the fitted regression models also for inference enables the consistent mapping of any contextualized BERT embedding to a moral vector.

Figure 1 visualizes the architecture of our approach. It also depicts which components are mainly responsible for context-awareness as well as generalizability/extendability to words currently not within the eMFD.

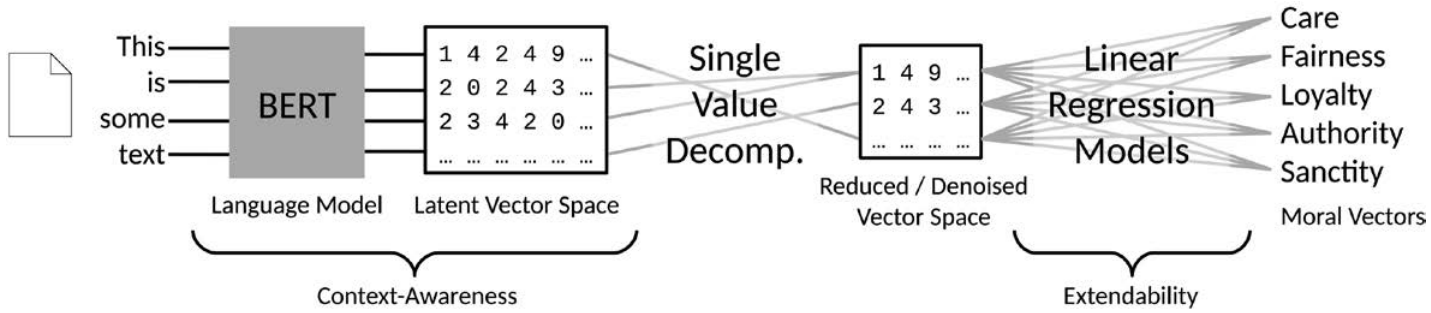


Figure 1: Model Architecture

BERT Word Embeddings
BERT allows for being fine-tuned by adding an additional layer, without having to train a full transformer model. Our model builds upon the hypothesis that the dimensions in BERT for representing words also encode moral connotations – even if not directly as moral vectors, but implicit as a linear combination of other vectors in the latent or reduced space. As we see in subsection 4.2, this hypothesis proves to be true. The implicit moral connotations are then extracted with multiple linear regressions using the ordinary least squares method [24].

To generate a context-aware embedding (i.e. vector) for each word of a text, we first frame the text between the [CLS] and [SEP] tokens. This instructs BERT to classify text [17]. The framed text is then tokenized using the pre-trained bert-base-uncased tokenizer1. The tokens are converted to token IDs and an attention mask considering all tokens is generated. Importantly, the BERT model can only handle a maximum of 512 tokens. Thus, we have to truncate the tokenized sequence to fit this restriction. Finally, the token IDs and attention mask are given to the pre-trained bert-base-uncased model. The BERT model is configured to return all hidden representations, so we can add the aforementioned fine-tuning layers.

Because the tokenizer might split individual words into multiple tokens so that each have a separate embedding, we have to aggregate these into a single embedding (i.e. vector) representing the combined word. This is achieved by averaging the moral scores of all tokens from each moral dimension, which together represent a single word. In other words, we take the average of a word's token embeddings to calculate the overall word embedding. As a result, the construction returns a list of tuples with each the word (as string) and a tensor, which contains a word's context-aware embedding (i.e. vector) for each of BERT's hidden states.

Validation
As we can see above, our approach effectively extends the eMFD beyond the range of its contained words. When comparing our model with the eMFD though, we have to ensure the same baseline with the eMFD. This means that we have to discard all words which are not contained in the eMFD from calculating the moral vector of the overall text. At the same time, we still would like to account for slight differences in moral connotations due to the words' context – which effectively is our improvement over the eMFD. To give an example, the moral vector of the following sentence is to be calculated: I love my parents. "I" and "my" may not occur in the eMFD – so we cannot compare our moral vector to the eMFD's vector for these words. However, due to exactly these words, there might be a slight difference in the moral vectors of "love" and "parents", because in our model these words are embedded in a different context. To account for this problem, we filter out words from calculating the moral vector without removing their context as follows.

As described in subsection 3.2, the words, for which one calculates the moral vectors, are first encoded into the embeddings – so they maintain their contextual meaning. It is important, that only thereafter the resulting tuples of words and their embedding vectors are filtered. This filtering mechanism allows for assessing only the words which also occur in the eMFD and comparing them using the aforementioned bag-of-words (BoW) method. The remaining vectors are then row-wise combined (i.e. stacked) to the embedding matrix X. By default, i.e. without this filtering, all words are considered as long as they are alphabetical (i.e. only contain letters) and are not stop-words.

Reproducing the eMFD
To assess our model and compare it to the eMFD vectors, we calculate the moral vectors from our model and the eMFD tool for all 1985 validation articles used in the eMFD paper [8] – the 1010 articles for developing the eMFD were excluded for selecting a proper baseline. This is similar to how the eMFD [8] was validated against other moral dictionaries, such as the original MFD [3] and the MFD 2.0 [27]. As there could have been some noise in our model due to the linear regressions, we wanted to see how well our model reproduces the eMFD. For proper comparison, we filter out all words not occurring in the eMFD as described before and use the BoW approach for calculating the overall moral vector for the article.

Generalizability beyond the eMFD
To assess the model's generalization capabilities, the eMFD is split into training and validation subsets, each representing 50%. The

[1] https://huggingface.co/google-bert/bert-base-uncased [17]

linear regressions of our model are inferred from the training subset. We then calculate the moral vectors for all words in the testing subset and compare them with the existing vectors from the eMFD. For this, the cosine similarity was found to be a proper metric [28].

Our model should not be restricted to the words in the eMFD, but should also provide reasonable moral scores for other words. We therefore assess whether our model's moral scores of words not contained in the eMFD follow the same distribution as our model's moral scores of words in the eMFD.

Short-Text Performance
As our model generalizes to other words not contained in the eMFD, we suspect a higher performance with short texts, because more words in short texts can be used for a moral analysis than with the eMFD. We therefore compare the moral vectors of the titles, summaries and full-texts of a random 10,000-article subset of the one million articles in the RealNews dataset [29] from our model with the ones from the eMFD. As ground truths for comparison, we calculate the moral vectors from the eMFD for the whole article. So, the closer the moral vectors from the short texts are to the moral vectors of the whole article, the better the model performs for short texts. As suggested in [29] (and later also seen in section 4.2), one cannot assume the meaning or moral vectors of headlines to align with the full-text article. However, our approach of benchmarking by the correlation of headlines to full-texts is still valid, because verifying a higher correlation is a stronger indication towards correct moral measurement than a lower correlation. This is supported by the fact that the eMFD cannot capture contextual semantics and thus yields inferior correlation.

Validation via Social Effects
However, it is not sufficient to validate our model against distributions. Technically, moral scores from our model can yield completely different moral scores while maintaining the exact same distribution as the eMFD. We thus compare our model identically to how the eMFD was compared against its predecessors [8]: analyzing the moral scores of articles from news sources with different political leanings. Previous research [3, 30] has found that conservative news sources tend to emphasize binding moral foundations (loyalty, authority, and sanctity) while liberal news sources tend to emphasize the individualizing moral foundations (care and fairness). This has been largely supported by the eMFD [8].

For comparing our model, we use the same three news sources with different political leanings: *Breitbart* [31] (far-right), *The New York Times* [32] (center-left), and *The Huffington Post* [33] (far-left). We then calculate the moral scores of articles from these news sources using our model and the eMFD.

Results

Reproducing the eMFD

Similar to the comparison of the eMFD with its predecessors [8], we compared statistical properties of our model's moral scores to the eMFD scores. The scores of both the eMFD and our model follow the same distribution (see Figure 2). Both are with some exceptions normally distributed, which aligns with the findings of the eMFD paper [8].

Interestingly, all of our models predict significantly higher probability scores for all moral dimensions compared to the eMFD.

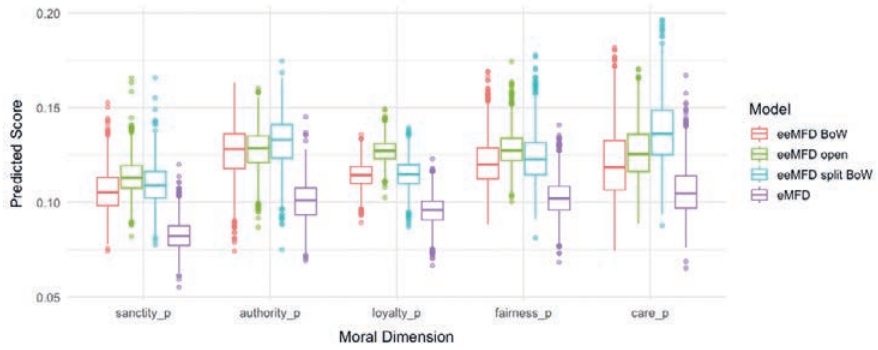


Figure 2: Distribution of Moral Scores of the Validation Articles between the eMFD and Our Model

We can also observe a visible shift in distribution within the moral dimension `care` for our model (see *eeMFD split BoW* in Figure 2). This is highly likely to be due to an unfortunate random training split, which biases the model to predict higher values. Analogously, the models *eeMFD BoW* and *eeMFD* open might not show this behavior as they were trained with the full training data, compensating for such bias. Besides the similar distribution of moral scores, we also analyzed their correlation (see Figure 3). Pearson correlation tests [26] within the same moral dimensions show that our model correlates with the eMFD scores between 0.71 and 0.82 (p-values 10^{-10}). Thus, we conclude that our model aligns with the existing moral understanding, as measured by the eMFD.

Generalizability

Our version of the model being trained only on the first half of the eMFD, also shows a correlation with the second half of the eMFD between Υ=0.59 and Υ=0.79 (*see eeMFD split BoW* in Figure 3). We conclude that our model effectively calculates moral scores of words which are not in the eMFD and can thus generalize. So, under the hood, our model finds relevant dimensions within the BERT language model that, in the main, strongly correlate[2] with moral scores in the eMFD. This is evidence for a very interesting insight about BERT's capability to recognize moral connotations.

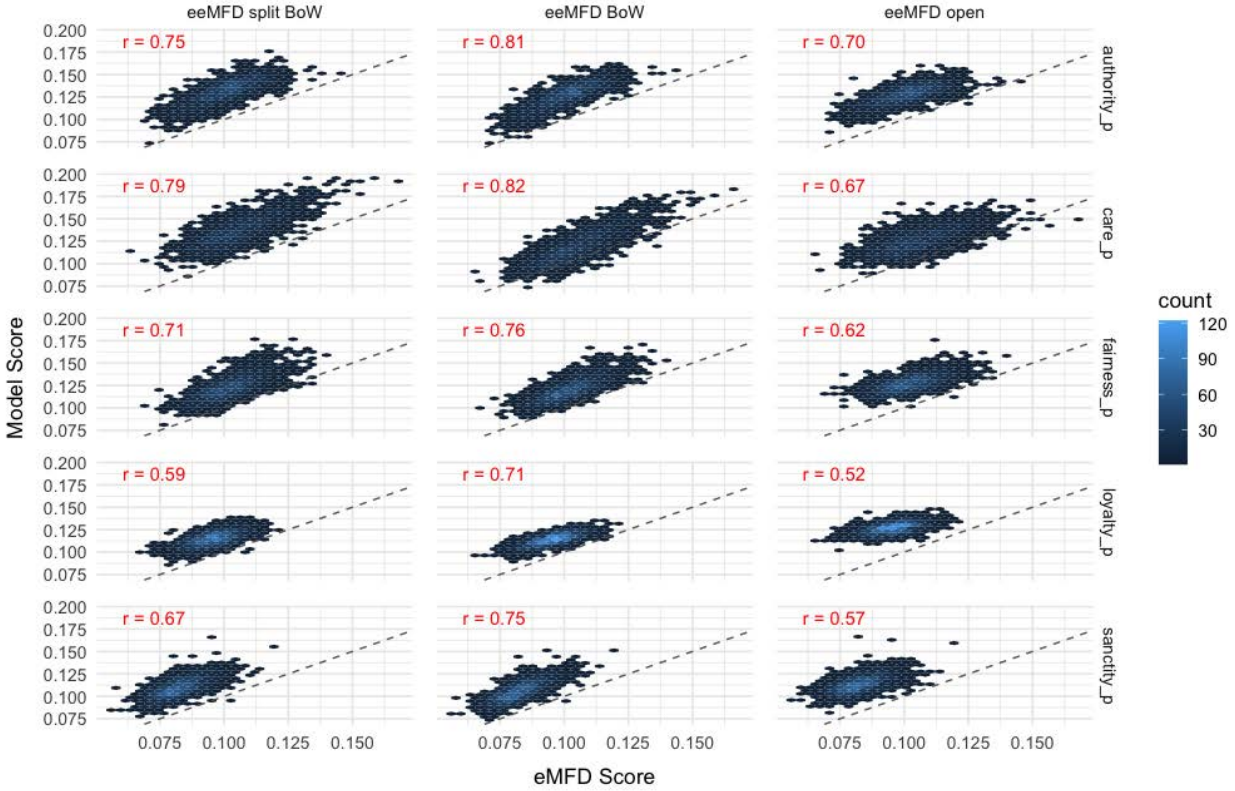


Figure 3: Rows represent the five moral dimensions. "eeMFD split BoW" shows the correlation results for testing generalizability of our model, "eeMFD BoW" shows the correlation results for reproducibility, and "eeMFD open" shows the correlation including words not within the eMFD. Each point represents the moral score of a single article in the validation dataset. The dashed diagonal represents a perfect correlation, i.e. a Pearson correlation Υ=1.

However, the correlations observed for our model trained on half the eMFD are lower compared to the model trained on the full eMFD. The difference can be explained in two ways. On the one hand, it may be that the smaller training dataset compared to a full eMFD makes generalization more difficult, simply due to its lower sample size. On the other hand, the model may overfit on the words within the eMFD. This would result in an inflated correlation optimized towards the training dataset, which does not properly reflect the model's generalization capabilities.

Our model should not be restricted to the words in the eMFD, but should also provide reasonable moral scores for other words. Fulfilling this requirement is supported by the fact that the moral scores follow the same distribution as the ones of words from the eMFD. For the correlation analysis in Figure 3, we therefore also compared a version of our model that excludes words in the eMFD (see eeMFD open Figure 3). Indeed, our model still shows a mostly strong correlation[2] with the scores of eMFD words. This indicates valid moral scores for any word in the BERT language model. Effectively, our model is a dynamic extension of the eMFD.

However, the aforementioned filtering for non-eMFD words is just for validation purposes. The full model (i.e. without word filters) yields a lower variance of the moral scores (see Figure 2) and correlates better with the eMFD than the word-filtered model for validation (see eeMFD open in Figure 3) – which is quite to be expected. Our full model yields lower absolute errors and fewer outliers than the eMFD (see Figure 4). This makes it a more reliable tool for short-text analysis.

Validation via Social Effects

The eMFD [8] was validated against previous moral dictionaries [3, 27] by analyzing the moral scores of articles from news sources with different political leanings. Previous research [3, 30] has found that conservative news sources tend to emphasize binding moral foundations (loyalty, authority, and sanctity) while liberal news sources tend to emphasize the individualizing moral foundations (care and fairness). This has been supported by the eMFD [8]. Having a look at Figure 5, the emphasis on binding moral foundations, especially authority and loyalty, are clearly visible for the far-right news source *Breitbart* in both models. However, the emphasis on

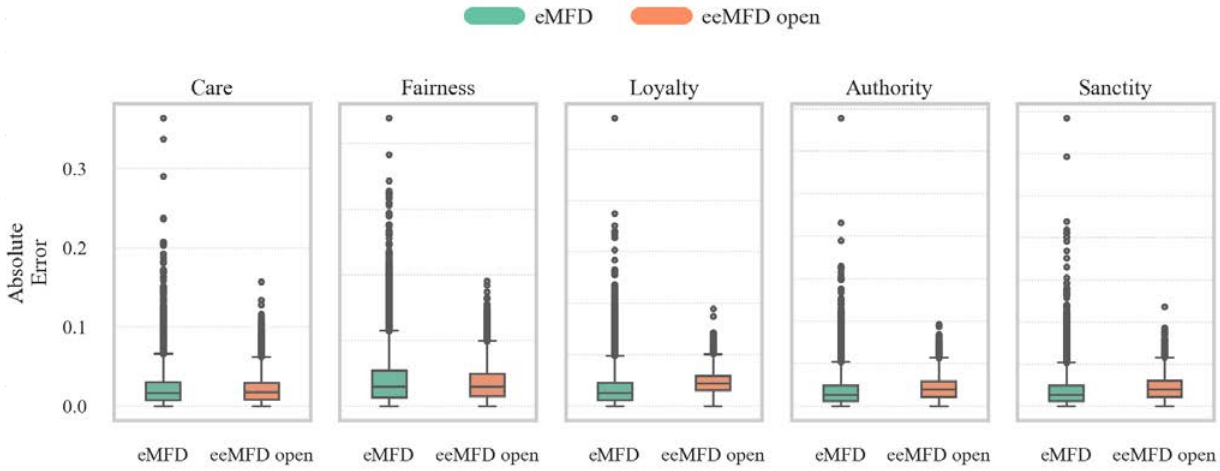


Figure 4: Absolute Error Distributions of the eMFD and Our Model

[2] Regarding a definition for which thresholds a correlation is weak, moderate or strong, we refer to the psychological definition, which is closest to our domain (see comparison in [34, Table 1]).

sanctity by *Breitbart* is not clear in either the eMFD or our model. Moreover, our model calculates higher care scores for *Breitbart* than for the other two news sources. This is in contrast to the eMFD, which shows *Breitbart* scoring similarly to *The New York Times* and *The Huffington Post.*

So overall, our model aligns with previous findings with some minor exceptions. This suggests that our model is a viable alternative to the eMFD.

Conclusion

In this research, we proposed an improved version of the eMFD which both includes contextuality to account for different meanings of the same words and shows improved performance for analyzing moral connotations on short texts.

We answered research question RQ1 regarding the context-awareness of moral measuring by combining the large-language model BERT with the eMFD. For this, we decomposed the moral vectors in the eMFD into separate vectors of the same words each in a different context (e.g. “love” as with parents or ice cream).

Research question RQ2 regarding the extension of the eMFD was answered with the same method, essentially focusing on the fact that BERT models many more words than the eMFD. By decomposing moral vectors onto the embeddings in the BERT model, we could calculate moral vectors for any word within BERT – including the ones not contained in the eMFD. Especially for calculating moral values of short texts, our model shows superior performance over the eMFD.

Our research is expected to impact multiple facets of computational social science – especially in the fields of detecting neutrality in text or fake-news detection in social media domains.

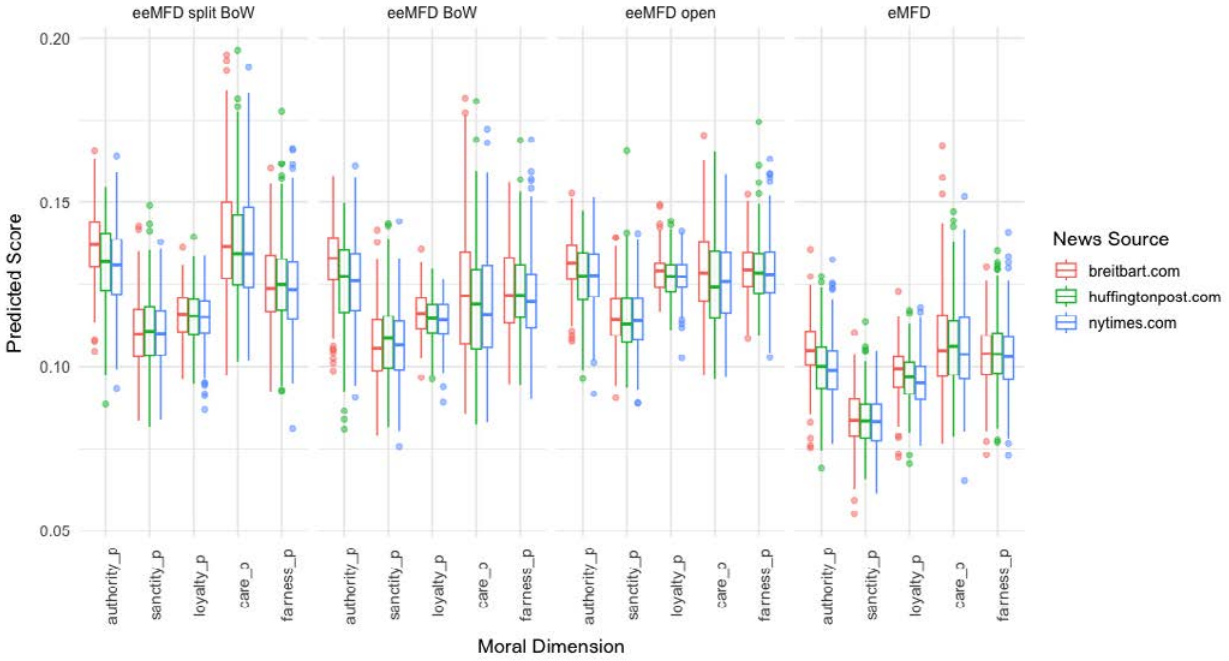


Figure 5: Moral Scores from the Validation Dataset Comparing Our Model to the eMFD

References

[1] [J. Haidt, The new synthesis in moral psychology. Science 316(5827), 998–1002 (2007). https://doi.org/10.1126/science.1137651

[2] [J. Grimmer, B.M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21(3), 267–297 (2013). https://doi.org/10.1093/pan/mps028

[3] [J. Graham, J. Haidt, B.A. Nosek, Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology 96(5), 1029 (2009). https://doi.org/10.1037/a0015141

[4] [J. Graham, J. Haidt. The moral foundations dictionary (2012). URL https://moralfoundations.org/wp-content/uploads/files/downloads/moral%20foundations%20dictionary.dic

[5] [R. Weber, J.M. Mangus, R. Huskey, F.R. Hopp, O. Amir, R. Swanson, A. Gordon, P. Khooshabeh, L. Hahn, R. Tamborini, Extracting latent moral information from text narratives: Relevance, challenges, and solutions. Communication Methods and Measures 12, 39–59 (2021). https://doi.org/10.1080/19312458.2018.1447656

[6] [J. Garten, J. Hoover, K.M. Johnson, R. Boghrati, C. Iskiwitch, M. Dehghani, Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior Research Methods 50, 344–361 (2018). https://doi.org/10.3758/s13428-017-0875-9

[7] [E. Sagi, M. Dehghani, Measuring moral rhetoric in text. Social science computer review 32(2), 132–144 (2014). https://doi.org/10.1177/0894439313506837

[8] [F.R. Hopp, J.T. Fisher, D. Cornell, R. Huskey, R. Weber, The extended moral foundations dictionary (eMFD): Development and applications of a crowdsourced approach to extracting moral intuitions from text. Behavior Research Methods 53, 232–246 (2020). https://doi.org/10.3758/s13428-020-01433-0

[9] [Y. Zhang, R. Jin, Z.H. Zhou, Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1, 43–52 (2010). https://doi.org/10.1007/s13042-010-0001-0

[10] E. Sagi, M. Dehghani, Moral Rhetoric in Twitter: A Case Study of the U.S. Federal Shutdown of 2013, in Proceedings of the Annual Meeting of the Cognitive Science Society (2014). URL https://escholarship.org/uc/item/9sw937kk

[11] [S. Roy, D. Goldwasser, Analysis of Nuanced Stances and Sentiment Towards Entities of US Politicians through the Lens of Moral Foundation Theory, in Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (2021). https://doi.org/10.18653/v1/2021.socialnlp-1.1

[12] [R. Rezapour, S.H. Shah, J. Diesner, 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 (2019). https://doi.org/10.18653/v1/W19-1305

[13] [R. Cao, M. Hu, J. Wei, B. Ihnaini, The Moral Foundations Weibo Corpus, in Proceedings of the 1st Workshop on NLP for Science (NLP4Science) (2024). https://doi.org/10.18653/v1/2024.nlp4science-1.13

[14] [M. Egorov, U. Steinberg, Moral Foundation Sensitivity: A Perspective Specific Moral Foundation Approach, in Academy of Management Annual Meeting Proceedings (2019), 1. https://doi.org/10.5465/AMBPP.2019.13307abstract

[15] [T.D. Nguyen, Z. Chen, N.G. Carroll, A. Tran, C. Klein, L. Xie, Measuring Moral Dimensions in Social Media with Mformer, in Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, vol. 18 (2024). https://doi.org/10.1609/20icwsm.v18i1.31378

[16] [N. Mokhberian, F.R. Hopp, B. Harandizadeh, F. Morstatter, K. Lerman, Noise Audits Improve Moral Foundation Classification, in 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2022). https://doi.org/10.1109/ASONAM55673.2022.10068681

[17] [J. Devlin, M.W. Chang, K. Lee, K. Toutanova, 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, vol. 1 (2019), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423. URL https://github.com/google-research/bert

[18] [N.M. Gardazi, A. Daud, M.K. Malik, A. Bukhari, T. Alsahfi, B. Alshemaimri, BERT applications in natural language processing: a review. Artificial Intelligence Review 58(166) (2025). https://doi.org/10.1007/s10462-025-11162-5

[19] [T. Mikolov, K. Chen, G.S. Corrado, J. Dean. Efficient estimation of word representations in vector space (2013). https://doi.org/10.48550/arXiv.1301.3781. International Conference on Learning Representations

[20] [A. Moschitti, B. Pang, W. Daelemans, GloVe: Global Vectors for Word Representation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014). https://doi.org/10.3115/v1/D14-1162

[21] [G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions. Numerische Mathematik (1970). https://doi.org/10.1007/BF02163027

[22] S.L. Brunton, J.N. Kutz, Data-Driven Science and Engineering (Cambridge University Press, 2022). https://doi.org/10.1017/9781009089517

[23] D.L. Donoho, M. Gavish, The optimal hard threshold for singular values is $4/\sqrt{3}$. IEEE Transactions on Information Theory 60 (2014). https://doi.org/10.1109/TIT.2014.2323359

[24] D.A. Freedman, Statistical Models: Theory and Practice (Cambridge University Press, 2009). https://doi.org/10.1017/CBO9781139165495

[25] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in Proceedings of the 14th international joint conference on Artificial intelligence, vol. 2 (1995). https://doi.org/10.5555/1643031.1643047. URL https://ijcai.org/Proceedings/95-2/Papers/016.pdf

[26] K. Pearson, Note on regression and inheritance in the case of two parents, in Proceedings of the Royal Society of London (1895). https://doi.org/10.1098/rspl.1895.0041

[27] J. Frimer, J. Haidt, J. Graham, M. Dehghani, R. Boghrati. Moral foundations dictionary for linguistic analyses 2.0 (2017). https://doi.org/10.17605/OSF.IO/EZN37

[28] W.P. Jones, G.W. Furnas, Pictures of relevance: A geometric analysis of similarity measures. Journal of the American Society for Information Science (1987). https://doi.org/10.1002/(SICI)1097-4571(198711)38:6⟨420::AID-ASI3⟩3.0.CO;2-S

[29] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending Against Neural Fake News, in Proceedings of the 33rd International Conference on Neural Information Processing Systems (Curran Associates Inc., 2019). https://doi.org/10.48550/arXiv.1905.12616

[30] D. Fulgoni, J. Carpenter, L. Ungar, D. Preot¸iuc-Pietro, An Empirical Exploration of Moral Foundations Theory in Partisan News Sources, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (2016), pp. 3730–3736. URL https://aclanthology.org/L16-159121

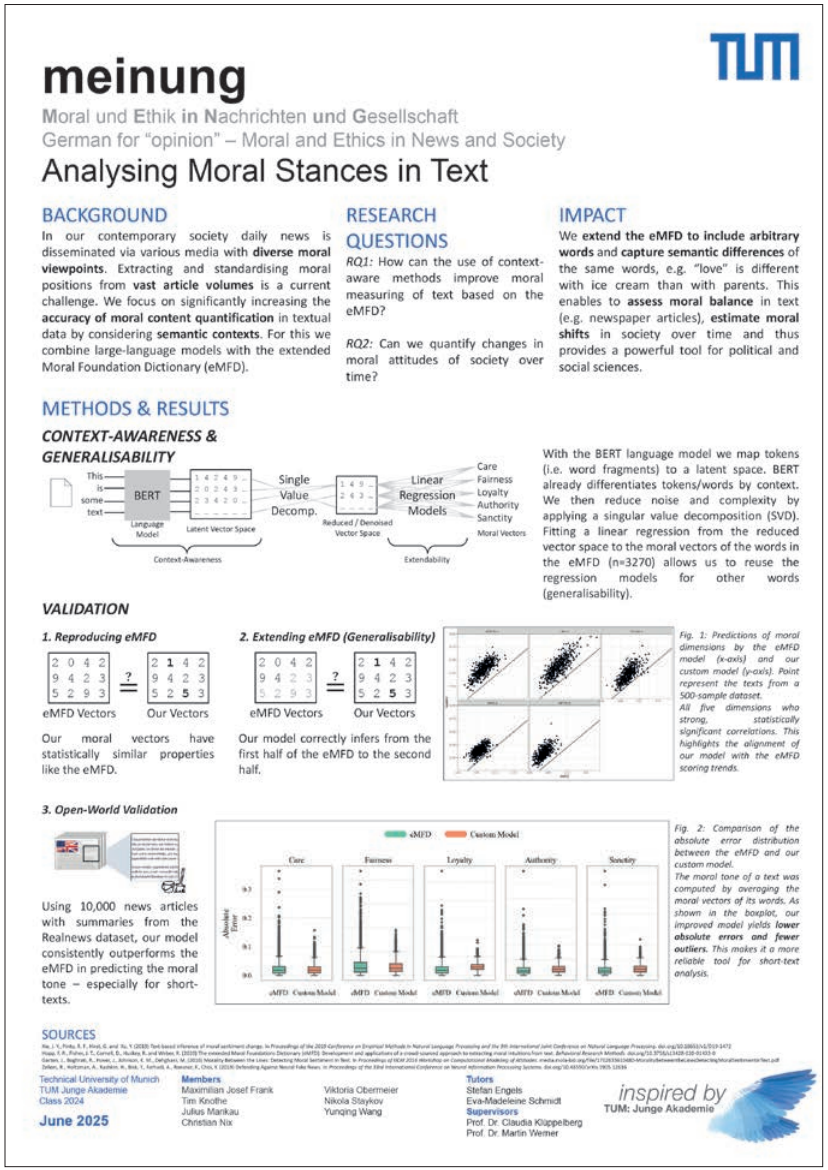
[31] Breitbart news network. URL https://breitbart.com

[32] New york times. URL https://nytimes.com

[33] Huffington post. URL https://huffpost.com

[34] H. Akoglu, User's guide to correlation coefficients. Turkish Journal of Emergency Medicine (2018). https://doi.org/10.1016/j.tjem.2018.08.001

Process description



Our interdisciplinary project began in late 2023 with the goal of analyzing moral language in newspaper articles. After an initial literature review, we structured our research around the Extended Moral Foundations Dictionary (eMFD), a tool that identifies moral content in texts. However, since the eMFD lacks context sensitivity – e.g., the word "love" varies by usage – we enhanced it with natural language processing (NLP) techniques to better capture contextual meaning.

We focused on various newspaper articles to explore two main questions: Can the eMFD be used to quantify shifts in moral attitudes during critical events? And how can its context sensitivity be improved using modern NLP methods?

Our analysis involved two parallel approaches: one purely dictionary-based, and one combining the eMFD with NLP techniques. This dual strategy enabled us to compare traditional and context-aware methods of moral analysis. We collected thousands of articles from US and UK news sources using a custom algorithm.

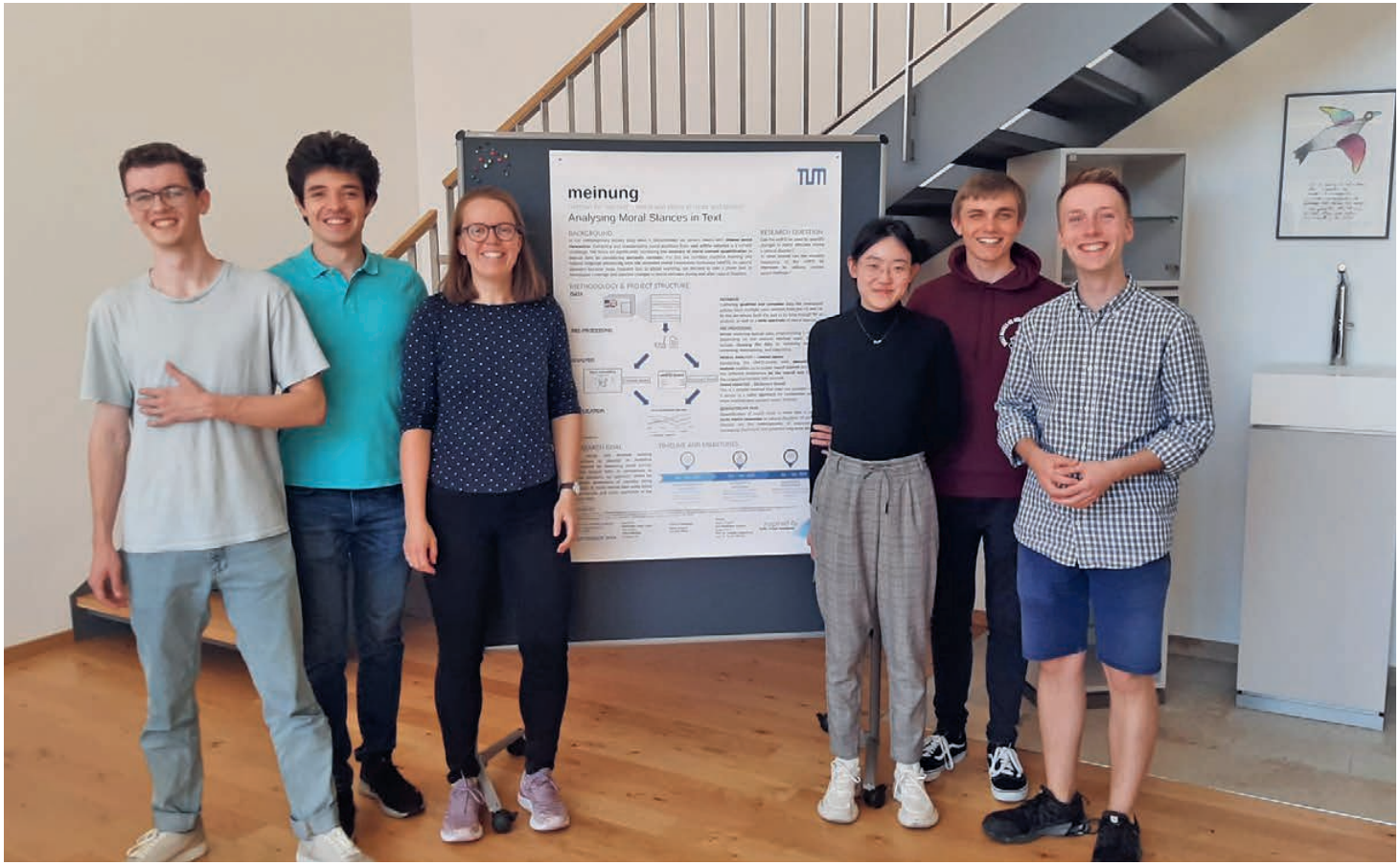
Throughout the project, we refined our methods during coding weekends, presented our results in a poster at a scientific conference, and incorporated feedback from both peers and experts. Our final phase included writing a research paper and preparing for the symposium, with a focus on tracing moral shifts over time and improving the eMFD's ability to capture nuanced moral expressions.

Self-Reflection

First and foremost, we would like to thank our supervisors, Prof. Claudia Klüppelberg and Prof. Martin Werner. Their consistent support and constructive feedback were instrumental in helping us stay focused and achieve results. We are also deeply grateful to our tutors, Stefan Engels and Eva-Madeleine Schmidt, who provided outstanding guidance throughout the project. Their availability and willingness to support us at any time ensured that we always had someone to turn to for advice.

Reflecting on the past 18 months, there may be some challenges leading to valuable insights and important lessons. These lessons can be crucial not only for personal development but also serve as a helpful guide for future teams at TUMJA.

Backup Plan: At the beginning, we needed to determine how to objectively measure morality, which seems to be impossible at first. After we realized the complexity of our task, we started to split



the project into two parallel sub-targets: first, continuing our initial goal to improve current methods of objective moral measuring (i.e., differentiating between words in different contexts by natural language processing), and second, measuring morality in a concrete application context. If we would have failed in our first goal, then we could still complete the second goal by using currently existing methodology (i.e., the pure dictionary approach). So, the extended Moral Foundations Dictionary (eMFD) not only became the foundation of our project but also served as our safety net at the same time. Even though we made progress in both goals, this strategy proved to be right due to unsatisfactory results. Additionally, narrowing down the topic to a concrete application gave us a clear vision of how our results might take shape. This eased the validation of the results regarding coherence and soundness.

Time-Management, Productivity, and Diversity: One of our biggest challenges was time management during peak periods of studying and working, especially exam periods. However, with our seven team members, we managed to distribute the workload quite well according to everyone's schedule and availability. Our team members came from diverse academic backgrounds, had varying strengths, and were often working in different countries.

This diversity required us to plan ahead while remaining flexible. We frequently adjusted the timing of our meetings, divided tasks, and allowed sub-teams to set their own schedules while adhering to overall deadlines. This structure allowed us to leverage each member's expertise and ensured that everyone could contribute meaningfully. In particular, the seminar weekends and our coding weekends, which combined project work with team-building, helped us stay focused and keep each other up to date. Especially after the initial project-forming phase, moving from weekly meetings to monthly coding weekends greatly increased productivity.

A valuable tip for future projects: Start working on your project early. This helped us immensely, especially when our initial ideas didn't stick.

Overall, the TUMJA research project was a great preparation for future scientific projects in interdisciplinary teams and for further specialization in our future careers. We value the connections we made – both professional and personal – and look forward to staying in touch.