Enhancing Stance Classification on Social Media Using Quantified Moral Foundations

Hong Zhang[1]*, Quoc-Nam Nguyen[1]*, Prasanta Bhattacharya[2], Wei Gao[1], Liang Ze Wong[2], Brandon Siyuan Loh[2], Joseph J. P. Simons[2], and Jisun An[3]

[1] School of Computing and Information Systems, Singapore Management University, 80 Stamford Rd, Singapore 178902
hong.zhang.2022@phdcs.smu.edu.sg, {qnnguyen, weigao}@smu.edu.sg

[2] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632
{prasanta_bhattacharya, wong_liang_ze, brandon_loh, simonsj}@ihpc.astar.edu.sg

[3] Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA
jisunan@iu.edu

Abstract. This study enhances stance detection on social media by incorporating deeper psychological attributes, specifically individuals' moral foundations. These theoretically-derived dimensions aim to provide an interpretable profile of an individual's moral concerns which, in recent work, has been linked to behaviour in a range of domains including society, politics, health, and the environment. In this paper, we investigate how moral foundation dimensions can contribute to detecting an individual's stance on a given target. Specifically, we incorporate moral foundation features extracted from text, along with semantic features, to classify stances at both message- and user-levels using traditional machine learning and Large Language Models (LLMs). Our preliminary results suggest that encoding moral foundations can enhance the performance of stance detection tasks, but with notable heterogeneity across task type, models, and datasets. In addition, we illustrate meaningful associations between specific moral foundations and online stances on target topics. The findings from this study highlight the importance of considering deeper psychological attributes in stance classification tasks, and underscore the role of moral foundations in guiding online social behavior.

Keywords: Stance Detection · Moral Foundations · User Behaviour · LLM

1 Introduction

As a behaviour, a *stance* refers broadly to an expression of perspectives, attitudes, or judgments toward a given proposition. The study of stances is inherently interdisciplinary and traces its roots to psycho- and socio-linguistics. In

[*] These authors contributed equally to this work.

Table 1: Sample tweets for the stance target *Hillary Clinton* and their expressed bias towards each moral foundation.

Stance	**Tweet**	**Care**	**Fairness**	**Loyalty**	**Authority**	**Sanctity**
FAVOR	@HillaryClinton the @DalaiLama speaks of women in leadership roles bringing about a more compassionate world. #potus #SemST	+	+	+	+	+
FAVOR	Just met an awesome supporter on the CX bus! He said "Hillary is one strong woman and we need that for our country." #FellowsNV #SemST	+	−	+	+	+
AGAINST	I wish #OliviaPope could run @HillaryClinton 's campaign... #Scandal #livisreal #SemST	+	−	+	−	−
AGAINST	Why did you lie about the #Benghazi subpoena? @HillaryClinton No wonder no one trusts you. #SemST	−	−	−	+	+
NONE	@larryelder The more Republicans talk about social policy the better....for #SemST	−	−	+	+	+

their popular text, Biber and Finegan [6] define stance as the “lexical and grammatical expression of attitudes, feelings, judgments, or commitment concerning the propositional content of a message”. Similarly, Du Bois [12] refers to stance as “an articulated form of social action”, which involves evaluation of an object, or an alignment with a given position.

Social media serves as a communicative platform that allows users to express their stance and views on a given topic of interest, providing researchers the opportunity to study this phenomenon at scale. Recent studies have proposed models for detecting or inferring the stance conveyed in social media posts on target topics [2, 24]. While the vast majority of these studies have leveraged message-level characteristics such as language use and user interactions, the question of whether stance modeling can be improved through the incorporation of deeper user-level attributes, notably psychological characteristics, remains understudied, with only a few recent exceptions [36].

Our working hypothesis is that a user’s broader value system carries relevant information for inferring their stance on a particular topic. In this study, we operationalize users’ broader value system along the conceptual framework of Moral Foundations Theory (MFT) [19, 22]. Notably, MFT proposes five distinct domains of moral concern rooted in universal evolutionary challenges: care/harm, fairness/cheating, authority/subversion, sanctity/degradation, and loyalty/betrayal. A key application case of this model is explaining political differences – liberals and conservatives endorse these moral foundations differently [18, 21].

Given the important role moral foundations play in shaping social behavior, we posit that they also enable the formation of human opinions and stances. In TABLE 1, we present a few tweets from the SemEval 2016 Task 6A dataset [28], which is widely used for stance detection tasks, and demonstrate that tweets across stance classes frequently express a positively- or negatively-valenced bias towards above-mentioned moral foundations. Hence, we hypothesize that the systematic extraction and incorporation of these moral foundations should enhance the detection of both, message- and user-level stances from social media-based content. In testing this hypothesis, we seek to augment online stance detection tasks on Twitter datasets by incorporating moral foundation features into message- and user-level stance detection models. Our findings reveal that the addition of moral foundation features significantly boosted the predictive performance of stance detection models. Furthermore, the insights generated from our association-based analyses highlight the prevalence and nature of moralized discourse surrounding key topics (e.g., wearing masks).

Through this study, we offer four key contributions. First, we perform a comprehensive analysis to assess the predictive utility of moral foundations for both message- and user-level stance detection tasks on social media. In doing so, we address the aforementioned limitations of current stance detection models, particularly their over-reliance on conventional textual and contextual features. Secondly, we highlight systematic variations in the predictive performance of moral foundation-based models across tasks (i.e., message-level vs. user-level stance detection), datasets, stance targets, and classifiers. Thirdly, we go beyond predictive performance to elucidate interesting associations between moral foundations and stances towards specific targets. Lastly, we show that the incorporation of moral foundations improves F1 scores on stance detection tasks by up to 24.86% points for LLM-based models, depending on the choice of model and dataset. This suggests that the addition of such psychological attributes might be particularly fruitful for LLM-based stance detection models. Further research can draw on these insights to explore the design of psychologically-rooted LLMs for related tasks.

2 Related Work

2.1 Stance Detection

As a natural language processing (NLP) task, stance detection has been widely studied in contexts spanning politics [39, 44], climate change [41], and the COVID-19 pandemic [17]. When expressed in text, the stance of the message can typically be labeled into a number of constituent classes, e.g., as “Support”, “Against” and “Neutral”. Stance detection aims to automatically determine the position of a message or its author towards a given proposition or target [28] based on these classes. In target-specific stance detection tasks, the models are trained for a particular target [4]. However, more recent work has investigated the problem of zero-shot stance detection which is target-agnostic and aims to detect the stance towards new or unseen targets [3].

2.2 Moral Foundation Dictionary (MFD)

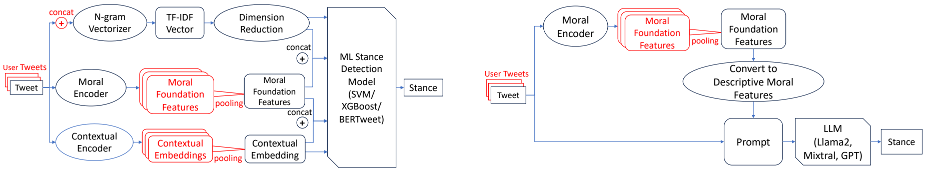
Moral Foundation Theory [20, 22] posits five moral foundations that are prevalent across cultures and nations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation. To measure these five dimensions from text, Graham et al. [18] developed the first MFD using a two-phase approach. The first phase involved generating associations, synonyms and antonyms for the five moral foundations through thesauruses and conversations with peers. Next, words that were too distantly related to the foundations were removed. This resulted in a dictionary of 295 words related to five moral foundations, which is also used in linguistic tools such as the Linguistic Inquiry and Word Count program (LIWC) [18]. Based on the MFD, Rezapour et al. [35] expanded the morality lexicons to a dictionary of 4,636 words, and used this to measure *social effects* such as morality and stances on Twitter.

2.3 Extended MFD (eMFD)

Although the MFD offers an automatic and dictionary-based method to extract moral foundation cues from text, it has a few limitations [14, 43]. For instance, the MFD was created by a small group of *experts* which limits its generalizability to a broader and *non-expert* population. Moreover, the MFD adopts a "winner takes all" strategy; it assigns a word to a single moral foundation. This precludes the possibility of a word being related to multiple moral foundations at once. To address these concerns, Hopp et al. [23] proposed the *extended* MFD (eMFD) to help capture large-scale and intuitive moral judgments from text. Instead of relying on careful selection by a small group of experts, the eMFD lexicon was generated through a wider crowd-sourced task aimed at capturing a more comprehensive list of morally relevant content cues. Moreover, instead of a single moral dimension, the eMFD assigns each word a vector of scores, which reflects the probability of that word belonging to each moral foundation.

2.4 FrameAxis

The MFD is only effective when the corpus of interest contains words from the dictionary. However, moral information in text can be expressed using diverse linguistic cues and styles, which might include only few or none of the words present in the MFD. In such contexts, it is challenging to infer the underlying moral foundations effectively using just a dictionary-based approach. In their study, Kwak et al. [25] proposed FrameAxis, a method for discovering the presence of framing and its associated biases from documents, by identifying the most relevant semantic facets. Using FrameAxis, any text can be projected to a high dimensional space using word embeddings to extract moral information, even when none of the words are present in the MFD.



(a) With traditional stance detection model.

(b) With LLM-based stance detection model.

Fig. 1: Our method for enhancing stance classification with quantified moral foundations based on traditional ML models and LLMs. Components in red were only used for user-level stance detection, where tweets posted by the same user were concatenated before passing through TF-IDF vectorizer, and pooling was applied for moral and contextual embeddings.

3 Datasets

We use three public datasets in our study, which we will subsequently refer to as SemEval[4], Connected Behaviour (CB)[5], and P-Stance[6] in this paper. The SemEval dataset was constructed for SemEval 2016 Task 6 [29], and contains 4,870 English tweets across six common targets, "Atheism" (AT), "Climate Change is a Real Concern" (CC), "Feminist Movement" (FM), "Hillary Clinton" (HC), "Legalization of Abortion" (LA), and "Donald Trump" (DT). The P-Stance [26] dataset is a popular stance detection dataset in the political domain, and contains 21,574 labeled tweets across 3 targets, namely "Donald Trump" (DT), "Joe Biden" (JB), and "Bernie Sanders" (BS). The CB dataset is a large Twitter dataset for *user-level* stance detection comprising over 100 million tweets [44]. This dataset was used for user-level stance detection towards three targets, namely "Donald Trump" (DT), "Wearing Mask" (WM), and "Racial Equality" (RE).

4 Our Method

In this section, we discuss how moral features were encoded and incorporated into both traditional ML models and LLMs. The performance gain in stance detection tasks with the addition of moral features was studied using the design shown in Fig. 1.

4.1 Feature Encoding

To incorporate moral features with stance detection models based on distinct learning methods, we used TF-IDF vectors, contextual embeddings, and moral

[4] `https://www.saifmohammad.com/WebPages/StanceDataset.htm`
[5] `https://github.com/tommyzhanghong/connected_behavior`
[6] `https://github.com/chuchun8/PStance`

foundation features. As shown in Fig. 1, tweets were passed through separate channels to generate these embeddings. Next, the dimensions of these embeddings were reduced using PCA and UMAP. These embeddings were then combined in different ways in the stance detection models. For the LLM-based models, in particular, the embeddings were converted to descriptive moral features, as shown in Fig. 1b and explained in Sec. 5.

TF-IDF. The term frequency-inverse document frequency (TF-IDF) vectorizer represents n-gram features by using both character- or word-level TF-IDF values. This has been applied in many studies spanning document retrieval, text classification and stance detection [13, 16, 33, 38].

Contextual Embedding. Sentence-BERT (SBERT) [34] is a variant of Transformer-based models [42] that uses siamese and triplet network structure in the training stage, and has been shown to outperform BERT and RoBERTa for sentence embedding tasks [27]. Here we encoded tweets with the SBERT all-mpnet-base-v2 version to generate contextual embeddings.

Moral Foundation Features. Two separate techniques were used to generate moral foundation features. The first was **eMFD** [23], a dictionary-based tool for detecting moral information in text. Using this method, *probability* and *sentiment* scores for a text along a moral foundation dimension were computed based on the frequency of occurrence of eMFD keywords in the text. This produced a 10-dimensional vector for each text (5 probabilities and 5 sentiment scores). The second was the Moral Foundation **FrameAxis** features [30], which combined the eMFD with the FrameAxis method described in [25]. In this method, the *bias* and *intensity* of a text along a moral foundation dimension are functions of the cosine similarity of word embeddings from the text and word embeddings of eMFD keywords for that moral dimension. This produced a 10-dimensional vector for each text (5 bias and 5 intensity scores).

User-level Representation. The CB dataset addresses the task of user-level stance detection. For this dataset, feature encodings were aggregated to form user-level representations. As TF-IDF is a statistical method relying on word and document frequencies, tweets posted by the same user were first concatenated to obtain a document for each of the targets. These user documents were then used to calculate TF-IDF embeddings. Conversely, other feature encodings have a fixed dimension. Hence we applied mean-pooling to obtain user-level embeddings from tweet embeddings.

4.2 Stance Detection Models

We compared three broad classes of models for the stance detection task.

1) Traditional Machine Learning Models In this study, we trained SVM and XGBoost classifiers using both n-gram and SBERT embeddings as basic features, and augmented these models by incorporating morality features, as described in the previous section. Such models have been widely used in stance detection studies [8, 28].

2) Fine-tuned Language Models Fine-tuned language models (FLMs) such as BERT-based models have been identified as state-of-the-art (SoTA) in stance detection tasks [5, 26]. In this study, we fine-tuned a pre-trained BERTweet model [31] on SemEval, CB and P-Stance datasets. BERTweet is the first public, large-scale and pre-trained language model for English tweets, having the same architecture as BERT-base [11]. Prior studies have shown that BERTweet outperforms strong pre-trained language models, and is SoTA on tweet-based stance tasks [10].

3) Large Language Models We implemented zero-shot and few-shot stance classification using three prominent LLMs: Llama2-70-chat[7], Mixtral-8x7B[8], and GPT-3.5-turbo[9]. Llama2-70-chat is based on the Llama 2 family of LLMs [40], and is fine-tuned for dialogue. Mixtral-8x7B is an LLM with a novel sparse mixture of experts (SMoE) architecture [37]. GPT-3 [7] is an autoregressive language model with 175 billion parameters, 10× more than any previous non-sparse language model. We chose these LLMs as they have performed well on many various NLP benchmarks, while also remaining cost-effective. Their pre-training data also includes social media posts, making them well-suited to process our tweet dataset. The LLMs were accessed via the Replicate API. Our prompting methods are presented in Section 5.

5 Prompting LLMs with Moral Features

5.1 Generating Descriptive Moral Features

As moral foundation features are represented as vectors of numbers, it is challenging for LLMs to interpret them. To address this challenge, we employed a two-step methodology: we first clustered these scores, and then converted the scores into textual expressions to be included in LLM prompts. Our method is described in Fig. 1b.

Discretization by Clustering: For each dimension of the eMFD or FrameAxis embeddings, we applied K-Means clustering with $K = 2$ on all scores for that dimension to determine the threshold between a *high* and a *low* score for that dimension. This allowed us to categorize each numerical score as either “High” or “Low”, which makes it easier for LLMs to understand the prompt.

Converting to Text: Based on whether each tweet was “High” or “Low” on each moral dimension, we generated textual moral descriptions for that tweet. These descriptions were then included in the stance detection prompts.

[7] `https://replicate.com/meta/Llama2-70b-chat`

[8] `https://replicate.com/mistralai/mixtral-8x7b-instruct-v0.1`

[9] `https://platform.openai.com/docs/models/gpt-3-5-turbo`

5.2 Prompting for Stance Detection

For stance detection with LLMs, we implemented three distinct prompting schemes. These schemes were designed hierarchically, with the second and third schemes built upon the information integrated in the first one.

Task + Context. The prompt included the task description and necessary contextual information, the stance target, and a clear definition of stance labels. This is consistent with previous studies leveraging ChatGPT for stance labeling [1].

Task + Context + FrameAxis. We further augmented the Task + Context prompt by including the moral descriptions generated from the FrameAxis embeddings.

Task + Context + eMFD. Alternatively, we augmented the Task + Context prompt by including the moral descriptions generated from the eMFD embeddings.

For the latter two schemes, we included the moral descriptions along with explanations of each moral dimension. Our hypothesis was that an LLM could make better stance predictions by considering these moral features. Our prompting strategies were adopted from the template samples available in HuggingFace's resources for LLaMa2[10] and Mixtral[11]. We implemented zero-shot, 1-shot, and 5-shot classification scenarios for the three schemes mentioned above.

6 Experiments and Results

6.1 Experimental Setup and Data Sampling

We tuned the hyperparameters for SVM and XGBoost following Gera and Neal [15]. We performed grid search on SVM with two kernels. For the Radial Basis Function (RBF) kernel, we used parameter values {1e-3, 1e-4} for *gamma* and {1, 10, 100, 1000} for *C*. For the linear kernel, we used parameter values {1, 10, 100, 1000} for *C*. Similarly, we performed hyperparameter tuning for XGBoost using grid search in {100, 500} for *n_ estimators*, {5, 10, 15, 20} for *max_ depth* and {0.01, 0.1} for *learning_ rate*. We applied a stratified 2-fold cross validation in the search process.

To correct for class imbalance in the SemEval dataset, we applied oversampling to balance training examples from different classes. In the CB dataset, we sampled 5 tweets from each of the top 500 active users, based on the number of posted tweets for each target and stance class, to avoid having an extremely large and sparse matrix due to the size of the vocabulary for TF-IDF. We did not observe a strong class imbalance for the P-Stance dataset, and hence, sampled 1,000 instances from the training data of each target for the same purpose.

We followed Nguyen et al. [31] to fine-tune BERTweet for each dataset and task over 30 epochs. We used AdamW with a learning rate of 1.e-5 and a batch

[10] `https://huggingface.co/blog/llama2`

[11] `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`

Table 2: Average % F1-scores across multiple datasets and models. Llama, Mixtral, and GPT denote Llama2-70b-chat, Mixtral-7x8B, and GPT-3.5-turbo, respectively. CB denotes the Connected Behavior dataset.

Features/Schemes	**Models**	**Datasets**		
		SemEval	**CB**	**P-Stance**
n-gram	SVM/XGB (+PCA)	50.21/46.52	81.99/78.88	67.14/66.00
n-gram + eMFD		51.93/48.09	82.55/80.05	67.85/67.74
n-gram + FrameAxis		52.49/47.53	82.19/80.35	68.98/68.79
n-gram	SVM/XGB (+UMAP)	42.03/44.87	79.65/76.33	66.02/66.02
n-gram + eMFD		44.20/47.70	80.57/77.32	66.81/66.81
n-gram + FrameAxis		45.21/47.80	80.60/78.09	67.37/67.37
SBERT	SVM/XGB	62.15/60.31	77.38/77.44	75.13/73.79
SBERT + eMFD		62.53/61.12	77.76/78.00	75.73/75.43
SBERT + FrameAxis		62.96/61.38	77.82/79.51	76.00/76.98
Tweet only	FLM (BERTweet)	71.26	65.50	79.29
Tweet + eMFD		71.69	66.31	80.68
Tweet + FrameAxis		75.30	70.07	79.88
Task + Context	Llama/Mixtral/GPT Zero-shot	54.14/37.19/66.46	43.48/22.01/72.28	61.94/22.57/74.80
Task + Context + eMFD		58.96/42.66/66.65	59.39/40.45/74.32	69.35/27.87/75.44
Task + Context + FrameAxis		58.61/39.35/66.82	63.70/34.37/74.00	80.58/34.65/75.48
Task + Context	Llama-/Mixtral/GPT 1-shot	55.48/39.18/68.57	44.03/24.77/73.71	62.82/23.90/77.88
Task + Context + eMFD		58.98/44.75/70.05	61.50/42.89/76.73	67.76/29.16/78.68
Task + Context + FrameAxis		59.50/47.52/71.41	67.71/44.83/78.14	79.95/39.30/80.91
Task + Context	Llama/Mixtral/GPT 5-shot	59.04/41.41/71.41	46.75/27.92/75.65	62.36/28.07/80.82
Task + Context + eMFD		63.99/47.86/74.39	68.14/50.37/79.25	68.31/31.97/82.61
Task + Context + FrameAxis		66.93/49.37/75.88	69.56/52.78/81.02	80.82/42.50/84.75

size of 32, and assessed performance after each epoch with early stopping applied if no improvement occurred over 5 epochs. We selected the best checkpoint for test set evaluation. For prompting-based experiments, we set the models to only provide the most probable output (e.g., by setting `temperature=0`) and a maximum length of 5 tokens [9, 32]. We repeated each experiment 10 times with random seeds, and reported the average F1 score as the final score from this exercise.

6.2 Results and Analysis

It is important to highlight that the goal of this study is not to produce a state-of-the-art stance detection model that can outperform existing stance detection benchmarks. Instead, we aim to study the effectiveness of encoding moral foundations in improving performance on stance detection tasks across a variety of models and datasets. We illustrate the experiment results in TABLE 2.

N-gram Baseline Using n-gram features as a baseline, the addition of moral features led to an improvement in the F1 scores on all datasets, with UMAP as a dimension reduction method showing greater improvement than PCA. The largest improvement with morality features was observed on the SemEval dataset, which has the smallest size. Specifically, the eMFD and FrameAxis features increased F1 scores by an average of 2.07% and 2.35% points, respectively, across models and dimension reduction methods.

SBERT Baseline To validate the effectiveness of moral foundation features with contextual embedding, the same experiments were repeated using SBERT embeddings [34] as the baseline model. Although contextual embeddings have been shown to be a stronger baseline than n-gram based models, we still observed improvements in stance detection performance on adding moral features. Notably, the addition of eMFD and FrameAxis increased F1 score by up to 0.59% and 0.94% points, respectively, for tweet-level stance detection on the SemEval dataset, and by up to 0.47% and 1.25% points for user-level stance detection on the CB datset.

Fine-tuned Language Models (FLMs) We conducted further experiments with FLMs, as they have achieved SoTA results in stance detection tasks [5, 26]. However, in our analyses, their performance was mixed, with significant variance across datasets. Specifically, we found that the FLM underperformed compared to our baseline model, SBERT, on the CB dataset, but showed superior performance on the SemEval and P-Stance datasets. Similar to the results observed with traditional machine learning methods, we observed a significant improvement in stance detection performance upon integrating moral information into the FLM. This integration led to improvements in F1 scores by up to 4.04%, 4.57%, and 1.39% points for SemEval, CB, and P-Stance datasets, respectively.

LLMs Prompting Our analysis revealed insightful trends and outcomes when evaluating the performance of Llama2-70b-chat, Mixtral-8x7B, and GPT-3.5-turbo models across three distinct learning scenarios: zero-shot, 1-shot, and 5-shot, and on the three datasets.

Zero-shot: As shown in TABLE 2, in the SemEval dataset, the inclusion of eMFD and FrameAxis led to performance improvements of up to 4.82% and 4.47% points, respectively, for Llama2-70b-chat, and up to 5.47% and 2.16% points for Mixtral-7x8B. In the CB dataset, enhancements were even more pronounced, with eMFD and FrameAxis boosting F1 scores by as much as 15.91% and 20.22% points for the Llama2-70b-chat model. For the P-Stance Dataset, improvements reached up to 7.41% points with eMFD and an impressive 18.64% points with FrameAxis, highlighting the substantial benefits of integrating moral foundation features.

1-shot: On the SemEval dataset, inclusion of eMFD and FrameAxis contributed to increases in F1 score by up to 3.50% and 4.02% points for Llama2-70b-chat, 5.57% and 8.34% points for Mixtral-7x8B, and 1.48% and 2.84% points for GPT-3.5-turbo. Similarly, for the CB dataset, including these moral features raised F1 scores by up to 17.47% and 23.68% points using the Llama2-70b-chat model, underscoring the value of incorporating moral foundation features. In the P-Stance dataset, the eMFD and FrameAxis moral features boosted F1 scores by 4.94% and 17.13% points using the same model, respectively, highlighting the effectiveness of FrameAxis in capturing moral narratives.

5-shot: On the SemEval dataset, we observed improvements of up to 4.95% and 7.89% points with the addition of eMFD and FrameAxis, respectively, for

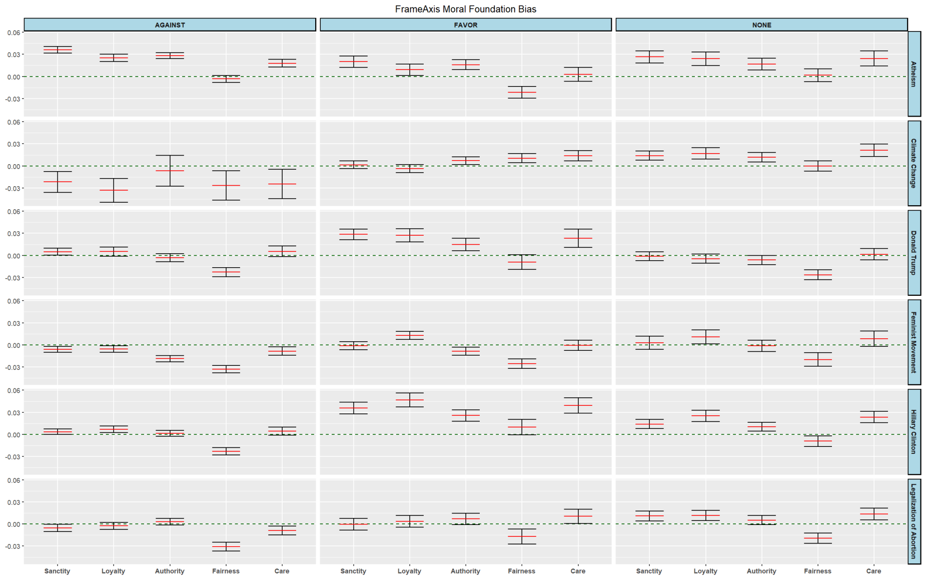


Fig. 2: Target- and stance-level heterogeneity in FrameAxis bias of moral foundations from the SemEval dataset.

the Llama2-7b-chat model. We noted similar improvements on the CB dataset using the Llama2-7b-chat model, with eMFD and FrameAxis leading to increases of up to 21.39% and 22.81% points respectively. Highest improvement was observed using Mixtral model, where the inclusion of FrameAxis led to performance improvement of up to 24.86%. In the P-Stance dataset, integrating eMFD and FrameAxis resulted in an increase of 5.95% and 18.46% points respectively, using the Llama2-7b-chat model. Taken together, these performance improvements highlight the predictive importance of morality features.

7 How do Moral Foundations Affect Stance Conveyance?

In the previous section, we highlighted the predictive value of incorporating moral foundations in stance detection tasks. In this section, we take a closer look at the associations between stances towards specific targets, and the expression of specific moral foundations.

In Figure 2, we illustrate the prevalence of various FrameAxis bias features for our focal moral foundations, across targets and stance classes in the SemEval dataset. We note that moral bias generated from FrameAxis varies significantly across targets, as well as between stance classes within targets. For example, most users against the target *Climate Change is a Real Concern* scored lower on the Care, Fairness, Authority and Sanctity foundations than users supporting the target. This was a result of users making greater use of language relating to

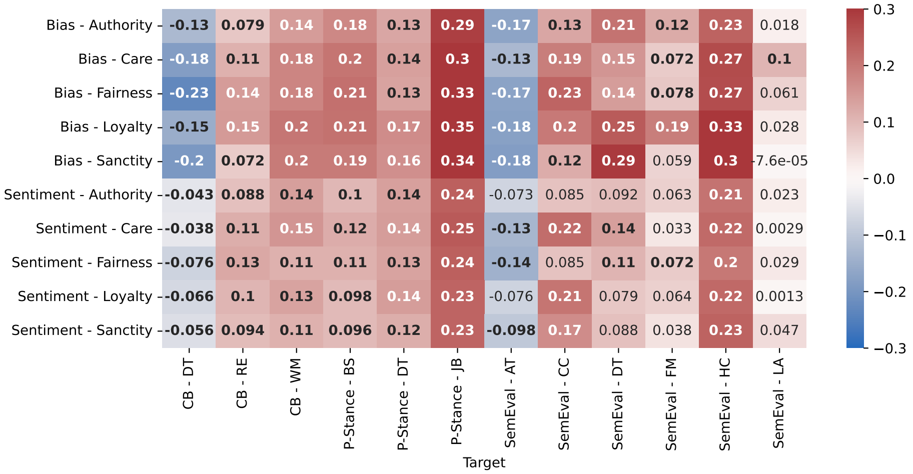


Fig. 3: Biserial correlations between moral foundation features and stance. Correlations that are statistically significant at $p < 0.05$ significance level are indicated in bold.

moral violation and/or lesser use of language relating to moral virtue, in these domains.

To further analyse how moral values are correlated with stances, in the SemEval and CB datasets, we encoded the "Favor" or "Support" stance as 1, and "Against" stance as 0, and subsequently measured the biserial correlation between this binary stance, and the sentiment and bias features for each moral foundation. The statistical significance based on a two-tailed t-test was calculated for each moral foundation feature and target pair.

As evident from Fig. 3, the eMFD Sentiment and FrameAxis Bias features exhibited the same direction of correlations for most targets and moral foundations. This reflects the consistency between the moral information extracted using eMFD and FrameAxis techniques. The direction of correlation provides an indication of the moral polarity between individuals supporting a target and those against it. The correlational results help explain why the elicitation of moral values from tweets might have contributed to the observed performance enhancements in our stance detection models.

8 Conclusion and Future Work

Moral foundations play a key role in shaping our social behavior in a variety of contexts. Although past studies have explored the prevalence and associations of moral foundations in online discourse, the predictive value of moral foundation representations across a diverse range of tasks, targets, and datasets have remained understudied. In this paper, we investigate if the inclusion of moral foundations can improve the detection of online stances on social media. While existing stance detection models primarily use text- and interaction-based features, our proposed models highlight the importance of incorporating deeper user-level attributes, such as their moral foundations. Our models show improved performance in both message- and user-level stance detection tasks using traditional machine learning models, FLMs and more recent LLMs. Additionally, we highlight insightful associations between stances and each of the five moral foundations, which can provide useful inputs for researchers studying online discourse around societal and political events.

The moral encoders used in this paper can also be improved using recent modeling innovations, notably the incorporation of LLMs. For instance, future work can consider using LLMs to generate moral foundation features, as well as other related attributes (e.g., personalities and beliefs) to further improve performance on stance detection and related tasks.

9 Acknowledgments

This research is supported by the SMU-A*STAR Joint Lab in Social and Human-Centered Computing (SMU grant no.: SAJL-2022-CSS02, SAJL-2022-CSS003). This research is supported by A*STAR (C232918004, C232918005).

Bibliography

[1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. Can we trust the evaluation on chatgpt? In *TrustNLP 2023*, pages 47–54, 2023.

[2] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[3] Emily Allaway and Kathleen Mckeown. Zero-shot stance detection: A dataset and model using generalized topic representations. In *EMNLP 2020*, pages 8913–8931.

[4] Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144, 2023.

[5] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *ACL findings: EMNLP 2020*, pages 1644–1650. ACL.

[6] Douglas Biber and Edward Finegan. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34, 1988.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[9] Iain J Cruickshank and Lynnette Hui Xian Ng. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734*, 2023.

[10] Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. Bernice: A multilingual pre-trained encoder for Twitter. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *EMNLP 2022*, pages 6191–6205. ACL.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL 2019: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL, June 2019.

[12] John W Du Bois. The stance triangle. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, 164(3):139–182, 2007.

[13] Johannes Fürnkranz. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10, 1998.

[14] Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. Dictionaries and distributions: Combining

expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods*, 50:344–361, 2018.

[15] Parush Gera and Tempestt Neal. A comparative analysis of stance detection approaches and datasets. In *Proc. Workshop on Eval4NLP 2022*, pages 58–69.

[16] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. Stance detection in fake news a combined feature representation. In *Proc. of Workshop on FEVER 2018*, pages 66–71.

[17] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. Stance detection in covid-19 tweets. In *Proc. of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Long Papers)*, volume 1, 2021.

[18] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029, 2009.

[19] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. volume 101, page 366. American Psychological Association, 2011.

[20] Jonathan Haidt. The new synthesis in moral psychology. *Science*, 316 (5827):998–1002, 2007.

[21] Jonathan Haidt and Jesse Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007.

[22] Jonathan Haidt and Craig Joseph. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66, 2004.

[23] Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53:232–246, 2021.

[24] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

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

[26] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. P-stance: A large dataset for stance detection in political domain. In *ACL findings: ACL-IJCNLP 2021*, pages 2355–2365.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[28] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on SemEval 2016*, pages 31–41.

[29] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM TOIT*, 17(3):1–23, 2017.

[30] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020*, pages 206–219. Springer.

[31] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In Qun Liu and David Schlangen, editors, *EMNLP 2020: System Demonstrations*, pages 9–14. ACL.

[32] Duc-Vu Nguyen and Quoc-Nam Nguyen. Evaluating the symbol binding ability of large language models for multiple-choice questions in vietnamese general education. In *SOICT 2023*, page 379–386.

[33] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proc. WMT 2015*, pages 392–395.

[34] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019*, pages 3982–3992.

[35] Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner. Enhancing the measurement of social effects by capturing morality. In *WASSA 2019*, pages 35–45, 2019.

[36] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics. In *Proc. ACM Hypertext 2021*, pages 177–188, 2021.

[37] Omar Sanseviero, Lewis Tunstall, Philipp Schmid, Sourab Mangrulkar, Younes Belkada, and Pedro Cuenca. Mixture of experts explained. Hugging Face Blog, 2023.

[38] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[39] Mariona Taulé, Francisco M Rangel Pardo, M Antònia Martí, and Paolo Rosso. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In *IberEval@ SEPLN*, pages 149–166, 2018.

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

[41] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. A multi-task model for sentiment aided stance detection of climate change tweets. In *AAAI ICWSM 2023*, volume 17, pages 854–865.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS 2017*, 30.

[43] René Weber, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. In *Computational Methods for Communication Science*, pages 39–59. Routledge, 2021.

[44] Hong Zhang, Haewoon Kwak, Wei Gao, and Jisun An. Wearing masks implies refuting trump?: Towards target-specific user stance prediction across events in covid-19 and us election 2020. In *ACM WEBSCI 2023*, pages 23–32.