Validation of Moral Contagion and Rumor Diffusion Theory in Online Social Networks

Introduction to the Study

The digital transformation of human communication has fundamentally altered how moral values propagate through society, with online social networks serving as lightning fast, and temporally salient conduits for the spread of moral discourse and influence. Moral contagion theory, which posits that morally-charged content exhibits increased virality compared to morally-neutral content, has emerged as a central framework for understanding the diffusion of political, emotional, and ethical debates in digital environments (Brady et al., 2017). This theoretical framework suggests that moral and emotional content triggers psychological mechanisms that increase sharing behavior, leading to accelerated spread through social network structures. The implications of this phenomenon extend far beyond academic research, influencing political discourse, social movements, and ethical and psychological communication strategies in an increasingly connected world.

Despite initial empirical support for moral contagion effects, recent investigations have raised fundamental questions about the robustness and generalizability of these findings. Burton, Cruz, and Hahn's critical reanalysis revealed potential methodological limitations and measurement artifacts that may have inflated estimates of moral contagion effects (Burton et al., 2021). These concerns highlight a broader challenge in computational social science: the sensitivity of theoretical conclusions to specific measurement approaches and analytical choices. The debate surrounding moral contagion exemplifies how methodological decisions - particularly those related to the operationalization of measuring the spread, longevity, and dispersion of moral content - can significantly influence empirical findings and theoretical interpretations within research on digital and online behavior.

The present paper addresses these methodological concerns through a comprehensive three-study design that systematically examines measurement bias, temporal and longevity dynamics, and the assessment of causal mechanisms underlying moral contagion effects. Study 1 focuses on the validity of moral foundations measurement by replicating baseline moral contagion findings using different computational approaches to moral content detection, ranging from traditional dictionary-based methods to state-of-the-art large language models. Study 2 extends the analysis to longitudinal contexts, examining the survival rate of how moral contagion effects unfold and persist over time by testing longitudinal robustness of the lifetime of tweets

with different levels of moral salience. Study 3 employs advanced causal inference techniques, specifically double machine learning approaches, to establish causal relationships between moral content and and the virality of moral and emotional information. Through varied, multi-method investigation, we aim to resolve ongoing theoretical debates while providing methodological guidance for future research in computational moral psychology and studies on political digital communication.

Literature Review

The moral contagion effect describes how exposure to moral expressions influences individuals' moral judgments and behaviors (Brady et al., 2017; Mooijman et al., 2018). Brady et al. (2017), in their seminal study, provided compelling evidence that moral-emotional language on social media platforms like Twitter could proliferate rapidly, shaping public discourse. Subsequent replications, however, have identified potential measurement biases linked to dictionary-based sentiment analysis, noting inherent limitations such as a lack of context sensitivity and semantic precision (Jaidka et al., 2020; Rheault et al., 2016).

Addressing measurement bias, recent literature suggests employing advanced computational approaches like machine learning, deep learning, and large language models (LLMs), to achieve more context-aware and nuanced detection of moral expressions. For instance, Garten et al. (2018) demonstrated the efficacy of supervised ML algorithms in accurately classifying moral foundations in text data compared to traditional dictionary methods. Similarly, Hoover, Johnson, et al. (2020) used recurrent neural networks (RNNs) and long-short term memory models (LSTMs), classes of deep-learning algorithms, showcasing superior performance in detecting context-dependent moral-emotional expressions. More recently, LLMs such as OpenAI's GPT models have exhibited high performance in capturing moral nuances in textual data, outperforming both dictionary-based and simpler ML techniques (Bommasani et al., 2021; Brown et al., 2020; Hendrycks et al., 2021; Jiang et al., 2022; Liscio et al., 2023; Sap et al., 2022).

Given these developments, our study rigorously examines the moral contagion effect on Twitter while systematically addressing measurement bias through a comparative framework, first benchmarking new tools and then implementing them to measure moral contagion - following previous studies (Burton et al., 2021). Our technique will make more accurate the tools used by computational moral foundations researchers to understand moral contagion. Specifically, our methodology will incorporate and compare dictionary-based sentiment analyses, traditional pretrained moral foundations classifiers (i.e., the `MFormer` and `moral-strength` Python packages) (Kennedy et al., 2020; S. M. Mohammad, 2021; Schuster & Paliwal, 1997; Yang et al., 2019), and state-of-the-art LLMs (e.g., OpenAI's GPT-3.5, GPT-4) (OpenAI, 2023;

Ouyang et al., 2022) to classify our corpus according to moral foundations criteria. Our dataset comprises extensive Twitter corpora previously collected and validated for moral-emotional content (Brady, Gantman, & Van Bavel, 2020; Hoover, Johnson, et al., 2020).

Methodologically, the ML approach will involve zero-shot pretrained classifiers already tuned using labeled moral sentiment data (Bojanowski et al., 2017; Mikolov, Chen, et al., 2013). While there are many deep learning models out there that leverage pre-trained word embeddings (e.g., GloVe, BERT embeddings) to capture sequential context and syntactic relationships in moral language (Devlin et al., 2019; Pennington et al., 2014) we use a state of-the-art LLM methodology, employing prompt engineering with GPT-based models to classify our moral data and generate contextually precise predictions of moral-emotional content (P. Liu et al., 2023; Wei et al., 2022). Our research question is as follows:

> Do advanced methods (ML classifiers and LLMs) for detecting moral foundations online content improve the measurement of morality and its spread through social networks?

We first measure moral foundations classification using our tools, benchmarking to ensure improvement over ground-truth human annotated data, as well as the dictionary based method, and then conduct statistical analyses using negative binomial regression models, allowing for comparisons of moral contagion across measurement methodologies (Cameron & Trivedi, 2013; Hilbe, 2011). We now formulate our hypothesis:

> We hypothesize that the results across the dictionary-based, ML, and LLM methodologies sequentially improve, reducing measurement bias compared to dictionary methods, and provide even more accurate analysis of the moral contagion phenomenon on digital platforms as the tools increase in power and complexity.

Study 1 Methods

In this study, we build on the foundational work of Brady et al. (2017) while focusing on reducing measurement bias in mapping Moral Foundations Theory to English-language tweets. Our approach is contextual and technical through both seeking to mitigate measurement bias and uncover shifts in moral sentiments and discourse online through cutting edge natural language processing techniques (Patton et al., 2020). By altering only the measurement of moral expressions, and employing advanced, state of the art LLM models, along with older pretrained transformers, we aim to address limitations observed in simpler dictionary-based approaches, used by, e.g.,

Burton et al. (2021). Building upon these established and emerging methodologies, we utilize moral-emotional dictionaries as a baseline, then implement traditional NLP pretrained transformer models, and then few-shot classification using large language models (LLMs), to understand moral-emotional sentiment, diffusion, and reach - hoping to improve the measurement of sentiment detection. Our baseline is based on negative binomial regression models that use dictionary methods and are common in the literature (Cameron & Trivedi, 2013; Hilbe, 2011; Neumann & Rhodes, 2024).

Conventional dictionary-based methods, prototypically reliant on static word lists, have been widely used to identify moral content in text (Brady et al., 2017). However, these approaches can introduce systematic biases, as they may overlook linguistic variability and contextual nuances that are captured better by newer, more advanced models (Taha et al., 2024). To improve detection accuracy and mitigate bias, we adopt supervised machine learning classifiers that have been trained on ground truth, human annotated moral foundations data, and which we then employ on Burton et al. (2021)'s twitter datasets - hoping for an improvement in detection accuracy. These algorithms are more dynamic than dictionaries and are well-suited for high-dimensional text data, which is especially valuable for processing tweets which can contain messy abbreviations, hashtags, and links and which exhibit heterogeneous and sometimes incoherent syntactic structures.

In parallel, we implement a few-shot classifier using more advanced transformer architecture.[1] These models are trained on dense vector representations extracted from contextual embeddings. For our purposes we use OpenAI's GPT architecture due to its wide use in the mainstream, its low (relative) cost, and its user friendly API. This strategy reduces reliance on rigid lexicons - unable to understand the textual context surrounding moral-emotional keywords - and enhances the robustness of model predictions when handling diverse online discourse (Pavan et al., 2023). Our aim here is to reduce measurement bias from the traditional bag-of-words (dictionary) methods and then replicate Burton et al. (2021)'s result with additional, more accurate and precise, insights reflective of reality.

Building on advancements in transformer-based architectures, and drawing baseline measurements from the implementation of static machine learning, supervised, pre-trained NLP algorithms like MFormer and moral-strength,[2] we further press measurement for accuracy and signal by augmenting our methodology with LLM classification using standard prompting techniques, querying OpenAI's ChatGPT API. Our use of transformer-based tools capture deeper semantic and moral nuances using self-attention mechanisms (Vaswani et al., 2017). Through prompt engineering, and

[1] An early example of which is BERT (Reimers & Gurevych, 2019), but, as is widely known, now there are others that significantly outperform BERT and even Gemini and Llama, such as Claude and ChatGPT.

[2] Both of which are algorithms which classify text according to moral foundations (see Araque and Iglesias (2019) and Nguyen et al. (2024).

contextual awareness (coded or "taught" to the model during training) these models are capable of discerning context-specific moral signals that might otherwise be lost in simpler or less context-aware methods. Recent studies suggest that LLMs, when carefully calibrated, can mitigate biases introduced by limited lexical resources - in measurement of moral foundations-based emotion expression specifically (Zangari, Greco, et al., 2025). Thus, introducing this technique enables us to show how measurements of the distance between moral foundations and emotional sentiment in tweets have been skewed by poorly fitting models, and that it is possible, with new AI tools, to measure this diffusion with greater precision and less bias.

This study provides a pathway to understanding the drawbacks of dictionary and even traditional ML-based methods applied to moral foundations sentiment in textual data as compared to large language model classification. We thus, first, benchmark the four techniques - dictionary, MFormer, moral-strength, and ChatGPT-4.1-nano - using a ground truth moral foundations text corpus that was annotated by human experts. This multileveled strategy allows us to test a variety of models leveraging everything from heritage NLP (dictionary-based; arguably obsolete in the LLM era but relatively accurate and still widely used) to self-attention transformer mechanisms that encode and decode strings of tokens to compute context-aware representations of input tokens within the corpus (Vaswani et al., 2017). Both MFormer and moral-strength, we anticipate, will be eclipsed by LLM-based measurements of the depth of moral sentiment online, and the speed and amplitude at which they travel.

While our primary results lie in measuring the effectiveness of neural network-based computational approaches, negative binomial regression models remain a crucial contribution (Cameron & Trivedi, 2013; Hilbe, 2011). To parameterize these models, following Burton et al. (2021) we estimate diffusion changes due to the presence of moral sentiment differing by the measurement techniques demarcated by our four approaches. Previous research on moral contagion (e.g., Brady, Crockett, and Van Bavel (2020) and Brady et al. (2017)) has demonstrated that social media data often exhibit overdispersion in count outcomes, such as the frequency of moral-emotional words. In other words, social sentiments are often thought to spread faster, and more efficiently than they actually may. We suspect that a major driver of this phenomeona is that moral sentiment is modeled to travel or disperse when its presence is not accurately measured, and so results are actually understated. Negative binomial regression, using dispersion as a dependent variable and measurement techniques as sequential IVs, is well-suited to address time-dependent, textual data structures, and provides a comparative baseline for the supervised ML, and LLM-based methods.

The purpose of our analysis is to advance our understanding of how moral foundations manifest in English-language tweets primarily through correcting and assessing potential measurement bias. Ultimately, this approach offers a lens through which to

assess moral contagion dynamics, building research capacity to validate use cases for computational, AI powered text analytics methods that are focused on moral emotions and their dispersion online (Hirschhäuser & Winter, 2024).

Application of Moral Foundations Classifications to Political Tweets

Building on our methodological framework, we now describe how we deploy the dictionary, moral-strength, MFormer, and LLM tools to analyze moral rhetoric in politically charged tweet datasets. In these datasets, narratives often express moral and political language and emotional contestation. Therefore the ML and LLM tools should enable a more accurate measurement of moral foundations compared to traditional lexicon-based approaches (see, e.g., Frimer et al. (2019) and Hopp et al. (2021a)).

In line with prior work demonstrating the robust cross-domain performance of transformer-based classifiers (e.g., Devlin et al. (2019) and Y. Liu et al. (2019)), tweets from the Twitter (X) datasets are preprocessed by standard tokenization and normalization procedures. These steps explicitly address tweet-specific noise - such as hashtags, abbreviations, and URLs - which could negatively impact clear classification results for social media data (S. Mohammad et al., 2016; Tausczik & Pennebaker, 2010). Each tweet is then called by the respective classifiers and a classification result obtained. The machine learning strategies - MFormer and moral-strength - build on advances in probabilistic labeling and ideally improve upon static lexicon-based methods (Graham et al., 2013; Haidt & Graham, 2007). The LLM based classifier - GPT-4.1-nano - a model with very high intelligence and world knowledge, should also improve on the baseline dictionary classifier.

Once each tweet is annotated with moral foundation probabilities, the labels are aggregated to reveal distinct patterns of moral emphasis across time and throughout changing political circumstances. The respective models are employed to capture the contextual features of moral language and are sequentially used as hyperparameter-like variables that aim to isolate moral foundations sentiment from general text-based signals. This dual use - a comparative analysis of model accuracy and diffusion prediction - is crucial for ensuring that our statistical analysis (following Burton et al. (2021)) accurately reflects measurements of the influence of moral rhetoric in political discourse.

Results

Analysis of Binary (Moral or Non-Moral) Content Classification Methods

The comparative evaluation of binary moral foundations classification methods on ground truth datasets reveals significant performance variations across different algorithmic approaches. Our benchmarking results demonstrate that all three ML-based models substantially outperform traditional lexicon-based approaches - the dictionary-based method. This performance advantage aligns with recent findings in the moral foundations literature, where transformer-based architectures (i.e., Nguyen et al. (2024)) have consistently demonstrated enhanced capability in capturing nuanced moral language compared to rule-based dictionary systems (Zangari, Greco, et al., 2025). The substantial improvement observed with moral-strength and MFormer, and robust accuracy of GPT-4.1-nano reflects the broader trend in natural language processing where large language models exhibit superior performance in complex classification tasks, particularly when fine-tuned on domain-specific datasets spanning diverse moral contexts (Araque & Iglesias, 2019).

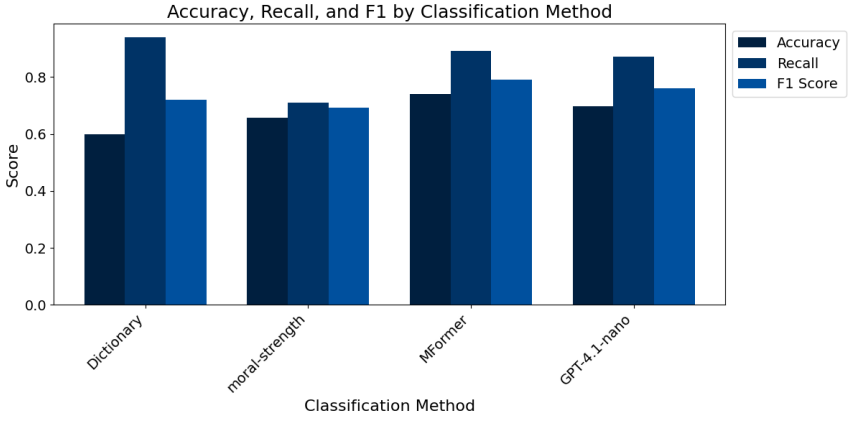

Figure 1: Accuracy, Recall, and F1 Score by Classification Method comparing Dictionary, moral-strength, MFormer, and GPT-4.1-nano approaches on the combined Reddit-MT and Twitter-MT datasets.

Moral Foundation Multi-Class Classification

Class-by-class analysis reveals distinct performance patterns across the six moral foundations, with particularly notable variations in recall and F1 scores for different moral dimensions (see Figure 2). The authority foundation demonstrates relatively

strong performance across all methods, achieving accuracy scores above 0.85 for both MFormer and GPT-4.1-nano, suggesting that authority-related moral language contains distinctive linguistic markers that facilitate accurate classification (Hopp et al., 2020). Conversely, the care and fairness foundations exhibit more challenging classification profiles, with lower recall scores indicating difficulty in identifying positive instances of these moral dimensions. This pattern is consistent with prior research highlighting the semantic complexity of care-based moral language, which often manifests through subtle contextual cues rather than explicit moral terminology (Crone et al., 2021). The purity foundation shows intermediate performance levels, reflecting the inherent complexity of disgust-based moral intuitions that often require sophisticated contextual understanding to identify accurately (Matsuo et al., 2019).

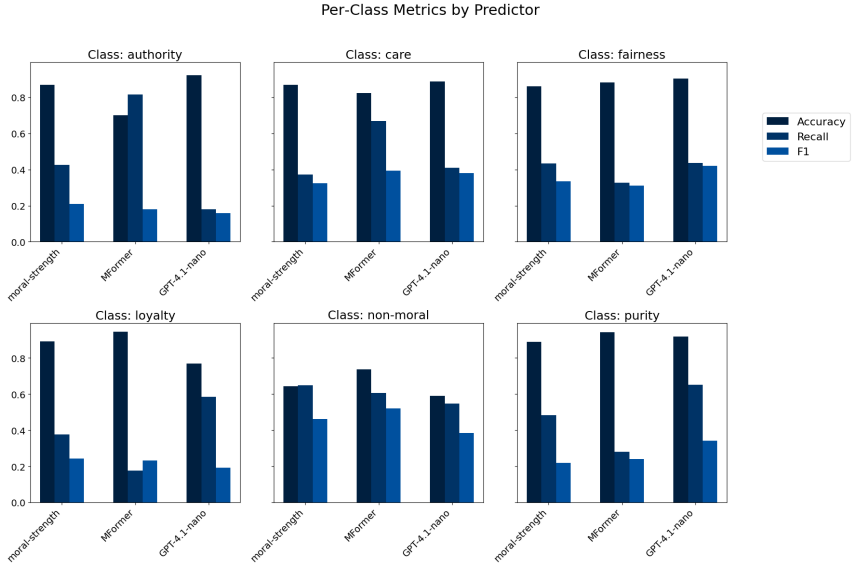

Figure 2: Per-Class Metrics by Predictor showing accuracy, recall, and F1 scores for each moral foundation (authority, care, fairness, loyalty, non-moral, purity) across moral-strength, MFormer, and GPT-4.1-nano methods.

Cross-Dataset Generalization and Model Robustness

The benchmarking results across Reddit-MT and Twitter-MT datasets illuminate important considerations regarding cross-platform generalization of moral foundations classification systems.[3] While all three statistical learning models consistently out-

[3]See Hoover, Johnson, et al. (2020) and Trager, Ziabari, Davani, et al. (2022) for explanations of these datasets.

perform baseline dictionary methods, the performance gaps vary significantly across different moral foundations classifications, suggesting that specific text-based attitude expressions, linguistic patterns, and context within the corpus influence the detection of moral expressions (Malik et al., 2021). The relatively lower performance of traditional dictionary-based approaches, particularly evident in the moral-strength baseline, underscores the limitations of lexicon-based methods when confronted with the dynamic and context-dependent nature of moral language in social media environments (Harris et al., 2022). These findings support the growing consensus in the moral foundations literature that effective moral classification requires sophisticated understanding of linguistic context, cultural nuances, and platform-specific communication patterns that extend beyond simple word-matching approaches (Van Vliet, 2021).

Results from Analysis of Canonical Models

Binary Classification Performance

The binary classification experiments aimed to determine whether social media text contains moral content, regardless of the specific moral foundation expressed. This fundamental task serves as a prerequisite for more granular moral foundations analysis and provides insights into the detectability of moral language across different platforms and methodological approaches, a common first step in MFT pipelines building on lexicons and annotated corpora (e.g., Araque et al. (2020), Graham et al. (2011), and Hopp et al. (2021b)).

Reddit Binary • Binary Classification

Method	Accuracy	F1 Score	Recall	N
Mf Binary	**0.716**	**0.704**	**0.716**	**32,935**
Gpt Binary	0.678	0.662	0.678	32,935
Mfd2 Binary	0.632	0.626	0.632	32,935
Ms Binary	0.619	0.616	0.619	32,935
Mfd1.0 Binary	0.617	0.618	0.617	32,935

Best: Mf Binary (0.716)

Figure 3: Performance comparison of binary classification methods on Reddit data. The MF Binary classifier achieved the highest accuracy at 71.6%, with consistent performance across all metrics. Methods are arranged by decreasing accuracy, showing a clear performance hierarchy among approaches.

Analysis of the Reddit binary classification results reveals a clear performance hierarchy among the tested methods. The MFormer binary classifier emerges as the

dominant approach, achieving 71.6% accuracy with balanced F1 score (0.704) and recall (0.716) metrics. This balanced performance profile suggests robust detection capabilities without sacrificing precision for recall or vice versa, in line with transformer-based MFT modeling that explicitly targets moral signals (Nguyen et al., 2023). The GPT-4.1-nano binary model follows as the second-best performer with 67.8% accuracy, demonstrating that transformer-based architectures can effectively capture moral content, though not quite matching the specialized MFormer approach (Abdulhai et al., 2024; Preniqi et al., 2024). Notably, the performance degradation from MFormer to moral-strength (ms) and the dictionary-based methods (MFD2.0, and MFD1.0) is relatively gradual, with all maintaining accuracy above 61%. This suggests that even simpler dictionary-based approaches retain meaningful signal detection capabilities for moral content in Reddit discussions (Araque et al., 2020; Hopp et al., 2021b), though the 10-percentage point gap to the best performer indicates substantial room for improvement through more sophisticated modeling approaches (Frimer, 2019).

Twitter Binary • Binary Classification

Method	Accuracy	F1 Score	Recall	N
Mf Binary	**0.764**	**0.757**	**0.764**	**35,799**
Gpt Binary	0.717	0.706	0.717	35,799
Mfd2 Binary	0.700	0.691	0.700	35,799
Ms Binary	0.690	0.689	0.690	35,799
Mfd1.0 Binary	0.681	0.681	0.681	35,799

Best: Mf Binary (0.764)

Figure 4: Performance comparison of binary classification methods on Twitter data. The MF Binary classifier demonstrates superior performance with 76.4% accuracy, showing notable improvement over Reddit results. The platform-specific performance gain suggests Twitter's format may facilitate moral content detection.

The Twitter binary classification results demonstrate consistently higher performance across all methods compared to Reddit, with the MFormer binary classifier achieving 76.4% accuracy—a notable 4.8 percentage point improvement over its Reddit performance. This platform-specific enhancement extends across all tested methods, suggesting that Twitter's constrained format may concentrate moral signals, making them more readily detectable (see the Twitter MFT corpus in Hoover, Portillo-Wightman, et al. (2020)). The F1 scores closely track accuracy values, with MFormer binary achieving 0.757, indicating well-balanced precision and recall. The GPT Binary model maintains its position as the second-best performer at 71.7% accuracy, though the gap to MFormer widens slightly on Twitter compared to Reddit. Moral-strength

and the traditional dictionary methods (MFD2.0, and MFD1.0) cluster more tightly on Twitter, ranging from 68.1% to 70.0% accuracy, suggesting that the platform's brevity may reduce the advantage of more sophisticated approaches for basic moral content detection, a pattern also noted in cross-platform analyses of Reddit vs. Twitter moral discourse (Hoover, Portillo-Wightman, et al., 2020; Trager, Ziabari, Mostafazadeh Davani, et al., 2022). This compression of the performance range raises interesting questions about the relationship between text length, moral expression density, and the comparative advantage of complex versus simple detection methods (Araque et al., 2020).

Multi-class Classification Performance

The multi-class classification task presents a substantially more complex challenge, requiring models to distinguish between specific moral foundations rather than simply detecting the presence of moral content. This fine-grained classification is essential for understanding the specific moral dimensions activated in social discourse and reveals the limitations of current computational approaches when faced with the nuanced nature of moral language (for theory and labels see Graham et al. (2009)and for recent multi-class modeling, Nguyen et al. (2023) and Guo et al. (2023).

Reddit Multiple • Multi-class Classification

Method	Accuracy	F1 Score	Recall	N
Mf Standardized	**0.438**	**0.451**	**0.438**	**32,935**
Mfd1 Standardized	0.409	0.379	0.409	32,935
Mfd2 Standardized	0.368	0.382	0.368	32,935
Ms Standardized	0.351	0.362	0.351	32,935
Gpt Mft	0.308	0.342	0.308	32,935
Emfd Standardized	0.180	0.124	0.180	32,935
Moralbert Standardized	0.180	0.124	0.180	32,935

Best: Mf Standardized (0.438)

Figure 5: Multi-class classification results on Reddit data showing performance across seven different methods. MFormer achieves the best performance at 43.8% accuracy, though all methods show substantial degradation compared to binary classification, highlighting the challenge of distinguishing between specific moral foundations.

The Reddit multi-class classification results reveal the substantial complexity inherent in distinguishing between specific moral foundations, with even the best-

performing MFormer Standardized method achieving only 43.8% accuracy. This represents a dramatic 27.8 percentage point drop from binary classification performance, underscoring the challenge of fine-grained moral categorization (also observed when moving from moral vs. non-moral to specific-foundation labels; see Hopp et al. (2021b)). The F1 scores show interesting divergence from accuracy metrics, with MFormer achieving 0.451, suggesting some robustness despite the lower overall accuracy. The traditional dictionary-based methods (MFD1.0 and MFD2.0) show moderate degradation to 40.9% and 36.8% accuracy respectively, while maintaining relatively competitive F1 scores. Most striking is the catastrophic failure of the supposedly advanced models—eMFD and MoralBERT—both achieving only 18.0% accuracy with F1 scores of 0.124. This unexpected underperformance suggests potential overfitting to different domains or fundamental misalignment with social media moral expression patterns (Preniqi et al., 2024). The GPT-4.1-nano model occupies a middle ground at 30.8% accuracy, indicating that general-purpose language models, while not optimal, avoid the complete failure modes of the specialized but potentially overfit models (Abdulhai et al., 2024).

Twitter Multiple • Multi-class Classification

Method	Accuracy	F1 Score	Recall	N
Mfd1 Standardized	**0.548**	**0.542**	**0.548**	**35,799**
Mf Standardized	0.530	0.545	0.530	35,799
Mfd2 Standardized	0.520	0.535	0.520	35,799
Ms Standardized	0.518	0.525	0.518	35,799
Gpt Mft	0.461	0.487	0.461	35,799
Emfd Standardized	0.263	0.185	0.263	35,799
Moralbert Standardized	0.263	0.185	0.263	35,799

Best: Mfd1 Standardized (0.548)

Figure 6: Multi-class classification results on Twitter data. Mfd1 Standardized achieves the highest accuracy at 54.8%, with Twitter showing consistently better multi-class performance than Reddit across most methods. The performance ordering differs notably from Reddit, suggesting platform-specific optimization opportunities.

Twitter's multi-class classification results present a notably different performance profile, with MFD1.0 Standardized emerging as the top performer at 54.8% accuracy—a surprising reversal from Reddit where MFormer led. This 10.4 percentage point improvement over Reddit's best multi-class result suggests that Twitter's constrained format may actually facilitate differentiation between moral foundations, not just

detection of moral content generally (Hoover, Portillo-Wightman, et al., 2020). The top four methods (MFD1.0, MFormer, MFD2.0, and moral-strength (ms)) cluster tightly between 51.8% and 54.8% accuracy, indicating reduced differentiation among dictionary and feature-based approaches on Twitter. F1 scores remain relatively strong, with MFormer Standardized achieving the highest at 0.545 despite not having the best accuracy, suggesting superior precision-recall balance. The GPT-4.1-nano model shows marked improvement on Twitter (46.1% vs 30.8% on Reddit), indicating that transformer architectures may be particularly sensitive to platform-specific characteristics (see also cross-domain fusion results in Guo et al. (2023)). However, eMFD and MoralBERT continue their poor performance at 26.3% accuracy, though this represents a modest improvement over Reddit. This persistent underperformance across platforms strongly suggests fundamental issues with these models' approach to moral foundations classification in social media contexts, potentially stemming from training on formal text that differs substantially from social media discourse patterns (Araque et al., 2020; Frimer, 2019).

Comparative Summary and Platform Analysis

Dataset	Best Method	Accuracy	Sample Size
Reddit Binary	MF Binary	**0.716**	32,935
Twitter Binary	MF Binary	**0.764**	35,799
Reddit Multiple	MF Standardized	**0.438**	32,935
Twitter Multiple	MFD1 Standardized	**0.548**	35,799

Figure 7: Summary of best-performing methods across all experimental conditions. The table highlights the consistent superiority of the MFormer method for binary classification and the platform-specific variations in multi-class tasks, with Twitter showing generally superior performance across all conditions.

The summary table crystallizes several key findings from our comprehensive evaluation. Most notably, MFormer binary dominates both platforms in binary classification, achieving 71.6% accuracy on Reddit and an impressive 76.4% on Twitter, establishing it as the clear choice for moral content detection tasks; this aligns with recent transformer-based advances tailored to moral signals (Nguyen et al., 2023) and with renewed scrutiny of dictionary reliability (Rehbein et al., 2025). The multi-class results tell a more nuanced story, with platform-specific winners: MFormer performs best on Reddit (43.8%) while MFD1.0 excels on Twitter (54.8%). This divergence suggests that optimal model selection for multi-class moral foundations detection must con-

sider platform characteristics. The consistent sample sizes (32,935 for Reddit, 35,799 for Twitter) ensure that performance differences reflect genuine model and platform effects rather than data volume disparities.

The platform-level performance gap is particularly striking: Twitter shows superior classification performance across all tasks, with binary classification accuracy 4.8 percentage points higher and multi-class accuracy 10.4 percentage points higher than Reddit. This systematic advantage likely stems from Twitter's 280-character limit forcing more concentrated and explicit moral language expression, whereas Reddit's long-form discussions may dilute or obscure moral signals across extended prose (Hoover, Portillo-Wightman, et al., 2020). The dramatic performance drop from binary to multi-class tasks (approximately 28–33 percentage points on Reddit, 21–22 percentage points on Twitter) reveals a fundamental challenge in moral foundations computation: while detecting the presence of moral content is relatively tractable, distinguishing between specific moral foundations remains a largely unsolved problem (Guo et al., 2023). This suggests that moral foundations may not manifest as discrete, easily separable categories in natural language but rather as overlapping, context-dependent constructs that resist simple classification schemes; survey work similarly calls for richer representations beyond label taxonomies (Zangari et al., 2025). Future work must grapple with this inherent complexity, potentially moving beyond traditional classification paradigms toward models that can represent the multifaceted and often simultaneous activation of multiple moral foundations in real-world discourse (see also framing and media evidence in Mokhberian et al. (2020b)).

Results from Sampled GPT-5 Classification

GPT-5 vs Human Annotation Performance

Figure 8 presents the comparative performance of GPT-5 against human annotations for Reddit and Twitter datasets across four macro-averaged metrics. GPT-5 achieved an accuracy of 0.382 for Reddit and 0.510 for Twitter, indicating better alignment with human labels on Twitter. Macro-precision, recall, and F1-scores were consistently higher for Twitter (0.382, 0.395, 0.369) than for Reddit (0.273, 0.282, 0.264), suggesting that GPT-5's classification boundaries align more closely with human judgments in short-form Twitter content than in Reddit's often more context-rich posts, echoing broader findings on LLM moral representations (Abdulhai et al., 2024) and guidance on LLM use in behavioral science (Abdurahman et al., 2024); see also (Bulla et al., 2025).

Error Distribution by Moral Foundation

Figure 9 shows confusion matrices for Reddit and Twitter. For Reddit, the model most accurately identified *Care/Harm* (85 correct) and *Non-moral* (177 correct) cases,

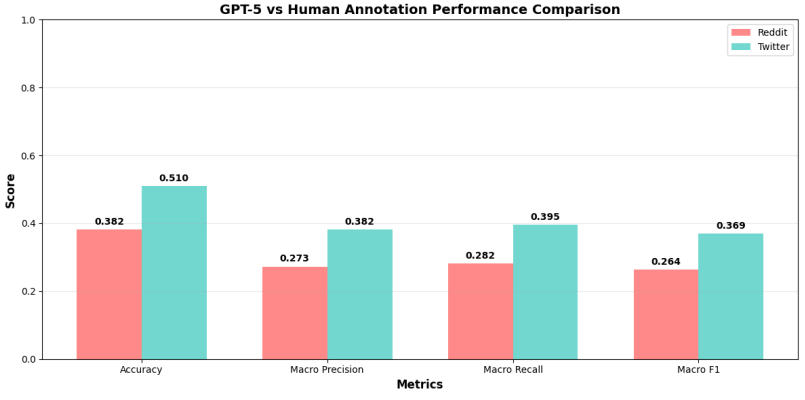

Figure 8: GPT-5 vs Human Annotation Performance Comparison.

but frequently confused *Fairness/Cheating* with *Care/Harm* and misclassified *Non-moral* content into moral categories. In Twitter, correct predictions were more evenly distributed, with strong performance on *Care/Harm* (135 correct) and *Non-moral* (200 correct), but notable leakage from *Non-moral* into *Care/Harm* and *Loyalty/Betrayal*; this pattern mirrors observed political tailoring and rationalization effects in LLM outputs (Simmons, 2023) and broader LLM moral benchmarking (Abdulhai et al., 2024).

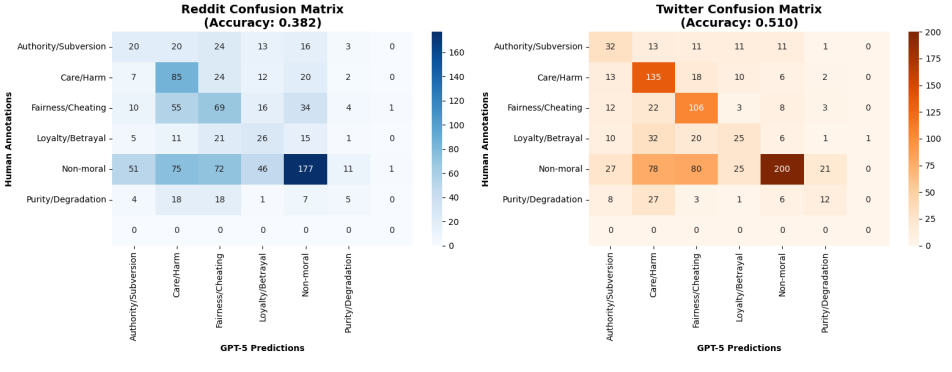

Figure 9: Confusion matrices for Reddit (left) and Twitter (right).

High-Confidence Error Analysis

A targeted review of high-confidence ($>$0.8) misclassifications revealed 37 cases for Reddit and 40 for Twitter. The most frequent error types involved misclassifying *Non-*

moral content as *Care/Harm*, followed by cross-moral misclassifications, particularly into *Care/Harm* from other categories such as *Loyalty/Betrayal* and *Purity/Degradation*, consistent with concerns about construct alignment and domain shift in moral text resources (Rehbein et al., 2025) and LLM moral reasoning variability (Abdulhai et al., 2024).

Confidence Calibration

Tables 1 and 2 summarize accuracy by confidence bin. On Reddit, GPT-5 exhibited substantial overconfidence below 0.8, with accuracy rates as low as 0.342 in the $0.5-0.7$ range. Twitter performance was more stable, with accuracy exceeding 0.74 in the $0.9-1.0$ bin; similar calibration issues have been reported when moral labels become fine-grained or cross-domain (Guo et al., 2023; Nguyen et al., 2023).

Table 1: Accuracy by confidence bin (Reddit).

Confidence Range	Accuracy	Cases
0.5–0.7	0.342	622
0.7–0.8	0.401	282
0.8–0.9	0.549	71
0.9–1.0	0.773	22

Table 2: Accuracy by confidence bin (Twitter).

Confidence Range	Accuracy	Cases
0.5–0.7	0.416	442
0.7–0.8	0.503	384
0.8–0.9	0.778	126
0.9–1.0	0.745	47

Justification Sentiment Analysis

Sentiment analysis of GPT-5's textual justifications indicates subtle valence differences between correct and incorrect classifications. In Reddit, correct cases had a slightly more negative mean compound score (-0.209) than incorrect ones (-0.191). On Twitter, correct justifications were also more negative on average (-0.080) than incorrect (-0.054), though the effect size was small. Across platforms, incorrect classifications

often had more neutral sentiment distributions, suggesting that strongly valenced justifications may correlate with higher model confidence, echoing findings that moral framing and affective tone shape interpretability and downstream judgments (Fantozzi et al., 2024; Mittal et al., 2023).

Summary, Key Insights, and Recommendations

Performance Insights

These results highlight first, that GPT-5 performs better on Twitter than Reddit for moral foundation classification, second, that the model exhibits systematic bias toward *Care/Harm* interpretations, especially at high confidence. Third, that confidence calibration is poor in mid-confidence ranges, particularly on Reddit, and finally, that sentiment polarity in justifications offers a potential auxiliary signal for error detection, which aligns with emerging evaluations of LLM moral bias and rationalization (Abdulhai et al., 2024; Simmons, 2023).

Overall, GPT-5 demonstrates higher alignment with human annotations on Twitter than on Reddit. The Reddit error rate is 61.8%, compared to 49.0% for Twitter, representing a 12.8% relative improvement on the Twitter dataset. This suggests that GPT-5 is more effective when handling shorter, more direct statements typical of Twitter, whereas Reddit's longer and context-rich discourse increases classification difficulty (Hoover, Portillo-Wightman, et al., 2020; Trager, Ziabari, Mostafazadeh Davani, et al., 2022).

Confidence Calibration

High-confidence errors—cases where GPT-5 is highly confident yet incorrect—represent a notable risk for downstream applications. For Reddit, 6.0% of all errors occur at high confidence, while for Twitter, this proportion is 8.2%. Although the absolute rates are modest, the fact that misclassifications can occur with such certainty highlights the need for calibration strategies, such as temperature scaling or confidence thresholding, before deployment; multi-domain fusion and dataset shift exacerbate these issues (Guo et al., 2023; Nguyen et al., 2023).

Common Error Patterns

Error pattern analysis reveals that *Authority/Subversion* is a frequent source of misclassification across platforms, often being re-interpreted as other moral foundations:

- **Reddit:** Most common confusions are into *Care/Harm* (20 cases), *Fairness/Cheating* (24 cases), and *Loyalty/Betrayal* (13 cases).

- **Twitter:** Similar trends, with *Authority/Subversion* most often reclassified as *Care/Harm* (13 cases), *Fairness/Cheating* (11 cases), and *Loyalty/Betrayal* (11 cases).

This indicates a systematic bias in GPT-5's internal mapping of *Authority/Subversion* toward other moral dimensions, potentially due to overlapping lexical or thematic cues and dictionary coverage limitations (Frimer, 2019; Hopp et al., 2021b), with cross-lingual dictionary work underscoring representation gaps (Cheng & Zhang, 2023).

Recommendations

First, adapt GPT-5 on moral foundations-specific training data to account for differences in discourse style and context density. Second, it might be possible to implement post-classification filtering that flags high-confidence disagreements for further review. Third, future research could introduce targeted contrastive learning examples, e.g., for *Authority/Subversion* vs. other foundations to reduce systematic confusion. Finally, a fruitful avenue might be to leverage justification sentiment and lexical focus to detect and correct bias within the weights of frontier and open-source models, combining transformer architectures with updated lexicons and task-specific supervision (Araque et al., 2022; Nguyen et al., 2023; Zangari et al., 2025).

Measurement bias

section intro

Women's March Twitter Dataset

Table 3: Effect of Moral Content on Tweet Diffusion (Retweet Count): Evidence from the Women's March Twitter Dataset

	Dependent variable:			
	Tweet Diffusion (retweet count)			
	(1)	(2)	(3)	(4)
Moral Tweet	−0.066	−0.107	0.037	0.047
	(0.144)	(0.144)	(0.113)	(0.113)
Emotional Language Use		0.238***	0.034	
		(0.077)	(0.060)	
Moral-Emotional Langauge Use			0.001***	0.001***
			(0.00001)	(0.00001)
Observations	3,654	3,654	3,654	3,654
Akaike Inf. Crit.	25,473.040	25,465.660	23,916.790	23,915.080
Note:				*p<0.1; **p<0.05; ***p<0.01

We find no effect of moral content on tweet diffusion on the Women's March Twitter Dataset, aligning with Burton (2021).

COVID-19 Twitter Dataset

Table 4: Effect of Moral Content on Tweet Diffusion (Retweet Count): Evidence from the COVID-19 Twitter Dataset

	Dependent variable:			
	Tweet Diffusion (retweet count)			
	(1)	(2)	(3)	(4)
Moral Tweet	1.750***	1.482***	1.208***	1.687***
	(0.262)	(0.257)	(0.258)	(0.264)
Emotional Language		1.364***	1.459***	
		(0.211)	(0.211)	
Moral-Emotional Language			0.775**	0.337
			(0.356)	(0.366)
Observations	994	994	994	994
Akaike Inf. Crit.	6,358.910	6,316.752	6,314.060	6,360.023
Note:				*p<0.1; **p<0.05; ***p<0.01

The LLM measurement of moral tweets is positively associated with diffusion and statistically significant at the 1% confidence level under all specifications. Results suggest that moral tweets receive, on average, 3.34–5.75 times more retweets than non-moral tweets (IRR 5.7 model 1–3.3 model 3).

Mueller Report Twitter Dataset

Table 5: Effect of Moral Content on Tweet Diffusion (Retweet Count): Evidence from the Mueller Report Twitter Dataset

	Dependent variable:			
	Tweet Diffusion (retweet count)			
	(1)	(2)	(3)	(4)
Moral Tweet	1.402***	1.384***	1.205**	1.238**
	(0.477)	(0.479)	(0.482)	(0.481)
Emotional Language		0.043	0.059	
		(0.239)	(0.239)	
Moral-Emotional Langauge			0.230	0.219
			(0.349)	(0.348)
Observations	995	995	995	995
Akaike Inf. Crit.	2,896.633	2,898.601	2,900.272	2,898.331
Note:				*p<0.1; **p<0.05; ***p<0.01

The LLM measurement of moral tweets is positively associated with diffusion and statistically significant at the 1% level under all specifications. Results suggest that moral tweets receive, on average, 3.3–4.1 times more retweets than non-moral tweets (IRR: 4.1 model 1–3.3 model 3).

MeToo Twitter Dataset

text here

Table 6: Effect of Moral Content on Tweet Diffusion (Retweet Count): Evidence from the MeToo Twitter Dataset

	Dependent variable:			
	Tweet Diffusion (retweet count)			
	(1)	(2)	(3)	(4)
Moral Tweet	0.421***	0.574***	0.653***	0.545***
	(0.126)	(0.126)	(0.127)	(0.127)
Emotional Language		−1.463***	−1.407***	
		(0.165)	(0.165)	
Moral-Emotional Language			−0.267**	−0.380***
			(0.127)	(0.129)
Observations	3,626	3,626	3,626	3,626
Akaike Inf. Crit.	12,787.030	12,732.240	12,730.300	12,780.950
Note:				*p<0.1; **p<0.05; ***p<0.01

The LLM measurement of moral tweets is positively associated with diffusion and statistically significant at the 1% level under all specifications. Results suggest that moral tweets receive, on average, 1.52–1.92 times more retweets than non-moral tweets (IRR: 1.52 model 1–1.92 model 3).

US Election Twitter Dataset

text here

Table 7: Negative Binomial Regressions

	Dependent variable:			
	Diffusion count			
	UMC	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Moral (LLM)	0.116** (0.049)	0.107** (0.049)	0.155*** (0.051)	0.165*** (0.051)
Moral–Emotional (recoded)		0.113*** (0.037)	0.111*** (0.037)	
ME_recoded			−0.207*** (0.054)	−0.209*** (0.054)
Constant	8.085*** (0.020)	8.022*** (0.028)	8.043*** (0.029)	8.105*** (0.021)
Observations	1,964	1,964	1,964	1,964
Akaike Inf. Crit.	35,579.730	35,572.410	35,560.330	35,567.430

Note: *p<0.1; **p<0.05; ***p<0.01

Post Brexit Twitter Dataset

Table 8: Negative Binomial Regressions

	Dependent variable:			
	Diffusion count			
	UMC	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Moral (LLM)	0.802*** (0.279)	0.883*** (0.279)	0.909*** (0.280)	0.848*** (0.280)
Moral–Emotional (recoded)		−0.292** (0.132)	−0.263** (0.131)	
ME_recoded			−0.450** (0.211)	−0.502** (0.211)
Constant	3.983*** (0.067)	4.134*** (0.100)	4.156*** (0.102)	4.024*** (0.070)
Observations	2,966	2,966	2,966	2,966
Akaike Inf. Crit.	15,544.120	15,541.240	15,539.190	15,541.120

Note: *p<0.1; **p<0.05; ***p<0.01

Brady et al. (2017) Results:

Brady et al. (2017) provide evidence that moral-emotional language systematically increases the diffusion of political messages on social media. Their analysis of tweets on gun control, same-sex marriage, and climate change revealed both similarities and differences in the effects of moral, emotional, and moral-emotional language.

For gun control, distinctly moral language and distinctly emotional language did not independently predict greater diffusion. In contrast, moral-emotional language exerted a clear positive effect: each additional moral-emotional word was associated with a 19% increase in expected retweets (IRR = 1.19, $p < .001$, 95% CI [1.14, 1.23]).

For same-sex marriage, moral language showed no effect (IRR = 0.99, $p = .540$, 95% CI [0.95, 1.03]), while emotional language significantly increased diffusion (IRR = 1.15, $p < .001$, 95% CI [1.11, 1.20]). Moral-emotional language also had a significant positive effect: each additional word predicted a 17% increase in expected retweets (IRR = 1.17, $p < .001$, 95% CI [1.09, 1.26]).

For climate change, all three factors were statistically significant predictors. Moral language had a modest but positive effect (IRR = 1.04, $p < .001$, 95% CI [1.02, 1.06]), emotional language was also significant (IRR = 1.08, $p < .001$, 95% CI [1.07, 1.09]), and moral-emotional language again produced the strongest effect, with each word increasing expected diffusion by 24% (IRR = 1.24, $p < .001$, 95% CI [1.22, 1.27]).

Measurement Bias: Summary Results

Table 9 compares our results to those reported by Burton, Cruz, and Hahn (2021). Whereas their dictionary-based approach yielded relatively modest and often non-significant effects of moral, emotional, or moral-emotional content on retweet counts, our LLM-based categorical coding consistently identifies stronger and statistically significant effects across issue domains. For high-salience issues such as COVID-19 and the Mueller Report, our results indicate much larger magnitudes of diffusion (IRR = 3.35 and IRR = 3.33, both $p < .001$) compared to Burton et al.'s more modest effects (IRR = 1.15 and IRR = 1.28, respectively). Similarly, we find stronger and now statistically significant positive effects for MeToo (IRR = 1.92, $p < .001$) and Post-Brexit tweets (IRR = 2.48, $p < .001$), in contrast to Burton et al., who found no positive effects in these cases (IRR = 0.91 and IRR = 1.02, respectively). Taken together, these comparisons suggest that LLM-based classifications capture the retweet advantage of moral, emotional,

and moral-emotional content more consistently and with greater effect sizes than dictionary-based word counts.

Our estimated effects are ∼3.3 times larger than Burton (2021) in the fully adjusted model and up to ∼5.75 times larger in the single-variable models.

Table 9: Results Comparison to Burton, Cruz, Hahn (2021)

Dataset	**IRR (Burton)**	**IRR**
Women's March	1.01	1.03
	(0.925)	(0.742)
COVID-19	1.15***	3.35***
	(<0.001)	(<0.001)
Mueller Report	1.28***	3.33***
	(<0.001)	(<0.001)
MeToo	0.91***	1.92***
	(<0.001)	(<0.001)
US Election	1.02	1.17***
	(0.465)	(<0.001)
Post Brexit	1.02	2.48***
	(0.370)	(<0.001)

Note: $^{*}p < 0.05, \quad ^{**}p < 0.01, \quad ^{***}p < 0.001$

Notes: Results compared using fully adjusted model (model 3 in regression tables), controlling for emotional and moral-emotional language.

Table 10: Results Comparison to Brady (2017)

Dataset	IRR (Brady)	IRR
Gun Control	0.98	3.73***
	(0.086)	(<0.001)
Same Sex Marriage	0.99	0.97
	(0.540)	(0.66)
Climate Change	1.04***	2.79***
	(<0.001)	(<0.001)

Note: $^{*}p < 0.05, \quad ^{**}p < 0.01, \quad ^{***}p < 0.001$

Notes: Results compared using single variable moral language factor.

Table 10 compares our findings to those of Brady et al. (2017). While their dictionary-based approach found only modest or null effects of moral content, our LLM-based categorisation identifies stronger and more consistent associations with retweet diffusion. For gun control, Brady et al. reported no significant effect of moral contagion (IRR = 0.98, p = .086), whereas our model shows a large and highly significant positive effect (IRR = 3.73, p < .001). For same-sex marriage, both approaches converge in finding no significant association (IRR = 0.99, p = .540 for Brady; IRR = 0.97, p = .66 for ours). For climate change, Brady et al. identified a small but significant effect (IRR = 1.04, p < .001), while our LLM-based classification reveals a substantially stronger relationship (IRR = 2.79, p < .001). Taken together, these comparisons suggest that LLM-based methods capture larger effect sizes than dictionary-based word counts.

Section Conclusion Overall, our findings suggest that prior studies may have underestimated the magnitude of the moral contagion effect. We provide evidence that moral content plays a larger role in information diffusion than previously recognised.

Implications Future Directions

The superior performance of transformer-based approaches, particularly MFormer and GPT-4.1-nano, has significant implications for the development of morally-aware AI systems and computational ethics applications. The consistent high accuracy scores (>0.85) achieved by these methods across multiple moral foundations suggest that large language models have reached sufficient sophistication to serve as reliable tools

for moral content analysis in research and practical applications (Priniski et al., 2021). However, the persistent performance variations across different moral foundations highlight the need for continued refinement in foundation-specific classification strategies, particularly for more subtle moral dimensions such as care and loyalty (Chen et al., 2021). Future research directions should focus on developing ensemble approaches that combine the semantic understanding capabilities of transformer models with the interpretability advantages of lexicon-based methods, while addressing the critical challenge of cultural and contextual bias in moral foundations detection across diverse social media platforms (Khoudary et al., 2022).

Discussion

The application of MFormer, moral-strength, and GPT-4.1-nano to these politically defined tweet datasets directly addresses a critical challenge in moral foundations analysis and classification: measurement bias. Traditional static word count methods (Graham et al., 2013) are prone to systematic errors that can understate subtle expressions of morality and exaggerate overt signals. By leveraging contextual embeddings, MFormer, moral-strength, and ChatGPT calibrate moral labels dynamically, thereby offering reduced measurement bias and capturing more accurately variation within moral rhetoric. This increased precision facilitates a more robust analysis of diffusion (see below) and the classification of ideological narratives, as shown by our results in conjunction with Burton et al. (2021) and Brady et al. (2017)'s original analysis (see also Mikolov, Sutskever, et al. (2013) and Reimers and Gurevych (2019)).

Moreover, integrating machine learning models like MFormer and moral-strength with established theories of moral reasoning (Haidt, 2012; Haidt & Joseph, 2004) incrementally helps to mitigate the limitations of earlier lexicon-based (dictionary) approaches. While we supposed that LLM models would also imporve on this, our findings did not support this conclusion.[4] As research on moral foundations expands to incorporate advanced NLP techniques (Mokhberian et al., 2020a), including LLMs, our results underscore how modern machine learning methods can overcome measurement bias, offering clearer insights into the relationship between language and social behavior - a promising development for both academic inquiry and practical applications in content moderation and political analysis.

[4]Note that larger and smarter models such as GPT-4o and o4 or Claude Opus were not tested as part of this study.

Conclusion

References

Abdulhai, M., Serapio-García, G., Crépy, C., Valter, D., Canny, J., & Jaques, N. (2024). Moral foundations of large language models. *Proceedings of EMNLP*. https://aclanthology.org/2024.emnlp-main.982.pdf

Abdurahman, S., et al. (2024). Perils and opportunities in using large language models in psychological science. *PNAS Nexus*, *3*(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Araque, O., Gatti, L., & Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations classification. *Knowledge-Based Systems*, *191*, 105212. https://doi.org/10.1016/j.knosys.2019.105212

Araque, O., Gatti, L., & Kalimeri, K. (2022). Libertymfd: A lexicon to assess the liberty moral foundation. *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (Companion)*. https://doi.org/10.1145/3524458.3547264

Araque, O., & Iglesias, C. A. (2019). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *arXiv preprint arXiv:1904.08314*. https://arxiv.org/abs/1904.08314

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). Moral contagion: Signatures and signals. *Current Opinion in Psychology*, *33*, 76–81.

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*(4), 746.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Bulla, L., et al. (2025). Large language models meet moral values [Online ahead of print]. *Patterns*. https://www.sciencedirect.com/science/article/pii/S2451958825000247

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Chen, K., Duan, Z., & Yang, S. (2021). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, 1–17.

Cheng, C. Y., & Zhang, W. (2023). C-mfd2.0: Developing a chinese moral foundation dictionary. *Computational Communication Research*, *5*(2), 1–47. https://doi.org/10.5117/CCR2023.2.10.CHEN

Crone, D. L., Rhee, J. J., & Laham, S. M. (2021). Developing brief versions of the moral foundations vignettes using a genetic algorithm-based approach. *Behavior Research Methods*, *53*(3), 1179–1187. https://doi.org/10.3758/s13428-020-01489-y

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fantozzi, P., et al. (2024). Detecting moral features in tv series with a transformer and mft. *Information*, *15*(3), 128. https://doi.org/10.3390/info15030128

Frimer, J., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). Moral foundations dictionary for linguistic analyses 2.0. *Unpublished Manuscript*, 1–6.

Frimer, J. (2019). Moral foundations dictionary 2.0. https://osf.io/ezn37/

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, *50*(1), 344–361.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (pp. 55–130, Vol. 47). Elsevier.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Guo, S., Mokhberian, N., & Lerman, K. (2023). A data fusion framework for multi-domain moral classification. *Proceedings of ICWSM*. https://ojs.aaai.org/index.php/ICWSM/article/view/22145/21924

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66.

Harris, C., Myers, A., & Kaiser, A. (2022). Being seen: How markets impact our moral sentiments. *Available at SSRN*. https://doi.org/10.2139/ssrn.3997378

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hirschhäuser, V., & Winter, S. (2024). Moral comments in social media: Analyzing the role of ideology-matching moral framing and impression motivation in the persuasive effect of user comments on youtube. *Mass Communication and Society*, *0*(0), 1–26.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Mostafazadeh Davani, A., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071. https://doi.org/10.1177/1948550619876629

Hopp, F. R., Fisher, J. T., & Weber, R. (2020). A graph-learning approach for detecting moral conflict in movie scripts. *Media and Communication*, *8*(3), 164–176.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021a). 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.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021b). The extended moral foundations dictionary (emfd): Development and applications. *Behavior Research Methods*, *53*(1), 232–246. https://doi.org/10.3758/s13428-020-01433-0

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165–10171.

Jiang, L., Hwang, J., Bhagavatula, C., Le Bras, R., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2022). Can machines learn morality? the delphi experiment.

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Khoudary, A., Hanna, E., O'Neill, K., Iyengar, V., Clifford, S., Cabeza, R., De Brigard, F., & Sinnott-Armstrong, W. (2022). A functional neuroimaging investigation of moral foundations theory. *Social Neuroscience*, *17*(6), 491–507. https://doi.org/10.1080/17470919.2022.2148737

Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C., Kalimeri, K., & Murukannaiah, P. (2023). What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, 14113–14132.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*, *55*(9), 1–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Malik, M., Hopp, F. R., Chen, Y., & Weber, R. (2021). Does regional variation in pathogen prevalence predict the moralization of language in covid-19 news? *Journal of Language and Social Psychology*.

Matsuo, A., Sasahara, K., Taguchi, Y., & Karasawa, M. (2019). Development and validation of the japanese moral foundations dictionary. *PLoS One*, *14*(3), e0213343. https://doi.org/10.1371/journal.pone.0213343

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mittal, S., et al. (2023). Moral framing of mental health discourse and its association with stigma. *ACM Transactions on the Web*, *17*(2). https://doi.org/10.1145/3544548.3580834

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41.

Mohammad, S. M. (2021). Ethics sheets for ai tasks. *arXiv preprint arXiv:2107.01183*.

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020a). Moral framing and ideological bias of news. *Social Informatics*, *12467*, 206–219.

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020b). Moral framing and ideological bias of news. *International Conference on Social Informatics (SocInfo)*. https://doi.org/10.1007/978-3-030-60975-7_16

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396.

Neumann, D., & Rhodes, N. (2024). Morality in social media: A scoping review. *New Media & Society*, *26*(2), 1096–1126.

Nguyen, T. D., Chen, Z., Carroll, N. G., Tran, A., Klein, C., & Xie, L. (2023). Measuring moral dimensions in social media with mformer.

Nguyen, T. D., Chen, Z., Carroll, N. G., Tran, A., Klein, C., & Xie, L. (2024). Measuring moral dimensions in social media with mformer. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 1134–1147.

OpenAI. (2023). Gpt-4 technical report.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Patton, D. U., Frey, W. R., McGregor, K. A., Lee, F.-T., McKeown, K., & Moss, E. (2020). Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 337–342.

Pavan, M. C., dos Santos, V. G., Lan, A. G. J., Martins, J. T., dos Santos, W. R., Deutsch, C., da Costa, P. B., Hsieh, F. C., & Paraboni, I. (2023). Morality classification in natural language text. *IEEE Transactions on Affective Computing*, *14*(1), 857–863.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Preniqi, V., Karaletsos, T., Bertagnolli, M., & Sadeghian, A. (2024). Moralbert: A fine-tuned language model for capturing moral values in social discourse.

Priniski, J. H., Mokhberian, N., Harandizadeh, B., Morstatter, F., Lerman, K., Lu, H., & Brantingham, P. J. (2021). Mapping moral valence of tweets following the killing of george floyd. *arXiv preprint*.

Rehbein, I., et al. (2025). Moral reckoning: How reliable are dictionary-based measures of morality in text? *Proceedings of NLP4DH*. https://aclanthology.org/2025.nlp4dh-1.20.pdf

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, *11*(12), e0168843.

Sap, M., Kiritchenko, S., Holgate, E., Mohammad, S. M., Bryant, K., Hovy, D., & Anastasopoulos, A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5886–5906.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Simmons, G. (2023). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *Proceedings of the 61st Annual Meeting of the ACL (Student Research Workshop)*, 282–297. https://aclanthology.org/2023.acl-srw.40.pdf

Taha, K., Yoo, P., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*, *54*, 100664.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Trager, J., Ziabari, A. S., Davani, A. M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., Reimer, N. K., Reyes, M., et al. (2022). The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Trager, J., Ziabari, A. S., Mostafazadeh Davani, A., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., Reimer, N. K., Reyes, M., et al. (2022). The moral foundations reddit corpus.

Van Vliet, L. (2021). Moral expressions in 280 characters or less: An analysis of politician tweets following the 2016 brexit referendum vote. *Frontiers in Big Data*, *4*, 49.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wei, J., Bosma, M., Zhao, V., et al. (2022). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Zangari, L., Greco, C. M., Picca, D., & Tagarelli, A. (2025). A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & Society*. https://doi.org/10.1007/s00146-025-02225-w

Zangari, L., et al. (2025). A survey on moral foundation theory and pre-trained language models [Online first]. *AI & Society*. https://link.springer.com/article/10.1007/s00146-025-02225-w

Zangari, L., Greco, C., Picca, D., & Tagarelli, A. (2025). A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & Society*, 1–26.