EMONA: Event-level Moral Opinions in News Articles

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, Ruihong Huang

Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{yuanyuan, huangrh}@tamu.edu

Abstract

Most previous research on moral frames has focused on social media short texts, little work has explored moral sentiment within news articles. In news articles, authors often express their opinions or political stance through moral judgment towards events, specifically whether the event is right or wrong according to social moral rules. This paper initiates a new task to understand moral opinions towards events in news articles. We have created a new dataset, **EMONA**[1], and annotated **e**vent-level **m**oral **o**pinions in **n**ews **a**rticles. This dataset consists of 400 news articles containing over 10k sentences and 45k events, among which 9,613 events received moral foundation labels. Extracting event morality is a challenging task, as moral judgment towards events can be very implicit. Baseline models were built for event moral identification and classification. In addition, we also conduct extrinsic evaluations to integrate event-level moral opinions into three downstream tasks. The statistical analysis and experiments show that moral opinions of events can serve as informative features for identifying ideological bias or subjective events.

1 Introduction

Morality refers to a set of social moral principles to distinguish between right and wrong (Berker, 2019; Nilsson et al., 2020). Moral judgment plays a crucial role in expressing public opinions, driving social movements, and shaping policy decisions. (Dehghani et al., 2016; Wolsko, 2017; Brady et al., 2020; Voelkel et al., 2022). *Moral Foundations Theory* (Haidt and Graham, 2007; Graham et al., 2009), provides a theoretical framework to categorize social moral principles into five dimensions, each associated with a positive and negative judgment: *Care/Harm*, *Fairness/Cheating*, *Loyalty/Betrayal*, *Authority/Subversion*, and *Purity/Degradation* (Figure 1 provides detailed values). Extracting moral framing from text demands a combination of sociological moral knowledge and contextual semantic understanding (Fulgoni et al., 2016; Xie et al., 2019; Johnson and Goldwasser, 2018).

In many studies, the moral foundations lexicon (Frimer, 2019) has been widely used to match words and identify moral foundations in text. Realizing the limitations of this lexicon match approach, researchers have started creating moral foundation annotations in text and training moral foundation detectors using annotations. However, such resource creation efforts have mainly been devoted to social media short text analysis and moral frames were usually annotated for an entire social media post (Trager et al., 2022; Roy et al., 2021; Hoover et al., 2020), in contrast, little work has explored moral sentiment within news articles at a more fine-grained and nuanced level yet.

In this paper, we propose a new task to understand moral opinions towards events in news articles. The concept of *event* refers to an occurrence or action, and is the basic element in story telling (Zhang et al., 2021; Lei and Huang, 2023b). In news media, the authors often express their stance through moral judgment towards events, so as to shape public opinions (Wolsko et al., 2016; Amin et al., 2017; Lei and Huang, 2022). To facilitate a profound study towards morality aspect of events, we create a new dataset, EMONA, annotated with Event-level Moral Opinions in News Articles.

While we believe the created dataset EMONA can be widely useful for studying moral opinion injection in context, this effort is initially motivated by the potential significant role of moral opinions for media bias analysis. In addition, recognizing the difficulty of identifying subjective events in news articles, we also aim to understand whether event moral opinions enables uncovering implicit subjective events that are otherwise hard to detect. To address these research questions, we have carefully chosen 400 documents from three sources

[1]https://github.com/yuanyuanlei-nlp/EMONA_dataset

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5239–5251

Care/Harm: should show care, generosity, compassion to others, sensitivity to suffering of others, and prohibit actions that harm others.

Fairness/Cheating: pursue fairness, justice, equality, and avoid discrimination, exploitation, or cheating.

Loyalty/Betrayal: should keep loyalty to in-group, virtues of patriotism and self-sacrifice for the collective, and punish betrayal.

Authority/Subversion: people should respect and obey authority and social traditions; authority should fulfill its social roles of maintaining social order.

Purity/Degradation: pursue spiritual purity, cleanliness, sanctity, and prohibit impure, disgusting, degrading actions.

Figure 1: *Moral Foundation Theory* categorizes moral principles into five dimensions.

Example 1: event-level moral opinions can be more implicit than opinionated content

In a succinct **speech [non-moral]**, the new president Donald Trump **told [non-moral]** Americans: "The time for empty **talk [cheating]** is over. Now arrives the hour of **action [authority]**."
The **speech [non-moral]** appealed for a "new national **pride [loyalty]**" and a "rediscovery of American **patriotism [loyalty]**".

Example 2: event-level moral opinions can reflect ideology bias

Left Ideology

Gov. Rick Perry of Texas is **broadcasting [non-moral]** a new campaign ad that seems to **contrast [cheating]** Christianity with homosexuality.
The Perry ad was **denounced [non-moral]** by the Log Cabin Republicans, a gay men and lesbian group, which **said [non-moral]** we should support **equality [fairness]** and our nation was **built [non-moral]** upon individual **liberty [fairness]**.

Right Ideology

Texas Gov. Ricky Perry is on the **attack [non-moral]**, **claiming [non-moral]** in a new TV ad that President Obama is waging a **war [degradation]** on religion, and he is the GOP candidate that can **defend [purity]** faith in America.
"I will **fight [purity]** against **attack [degradation]** on our religious heritage," **says [non-moral]** Perry in the new ad.

Figure 2: Examples of moral opinions towards events in EMONA dataset.

for annotation, including 180 documents from AllSides that spans over 12 domains and indicates article-level ideology bias (Baly et al., 2020); half (150 documents) of the BASIL dataset that has sentence-level media bias annotated (Fan et al., 2019); and the entire MPQA 3.0 dataset (70 documents) that has fine-grained opinions toward entities and events annotated (Deng and Wiebe, 2015).

The annotation process went through two passes: the first pass annotates event mentions, and the second pass annotates moral opinions for individual event mentions with respect to the context of the news article. Annotating event-level moral opinions within a news article turned out to be a challenging and demanding task for human annotators, we recruited five annotators and steadily improved pairwise inter-annotator agreements for both tasks to a satisfactory level. In total, the dataset contains over 10k sentences and 45k event mentions, among which, 9,613 event mentions were annotated as bearing moral opinions.

The challenge of event morality analysis indeed lies in the implicit nature of moral opinions toward events. In some cases, event-level moral opinions can be too implicit to be identified as opinions. Take the example 1 in Figure 2 as an instance, which are annotated as non-opinionated content by previous work (Fan et al., 2019; Deng and Wiebe, 2015), the semantic implies *cheating* criticism towards *empty talk*, and infers praise towards *patriotism* event. The implicit nature of event-level moral opinions requires the annotators to take both local sentential context and global broader context into consideration during annotations.

The numerical and visualization analysis of the dataset show that annotated event-level moral opinions can effectively reflect article-level ideology, and designate sentence-level political bias. Take the example 2 in Figure 2 as an illustration, liberal media focus on *fairness* judgment by framing the story as *Texas Gov. contrasts Christianity with homosexuality*, while conservative media emphasizes on *purity* moral by praising *Texas Gov. defends religious faith*. While not evident to readers, the journalists often subtly influence public opinions through moral value implication (Roy and Goldwasser, 2021). The annotated moral opinions towards events can uncover and provide fine-grained explanations for such ideological bias.

We build baseline models for event moral identification and classification, and we further conduct extrinsic evaluations on three downstream tasks: article-level ideology classification, sentence-level media bias identification, and event-level opinions identification. The experiments demonstrate the usefulness of detecting event-level moral opinions on all the three extrinsic evaluation tasks, with F1 score improved by 3.35% to 4.71%, and also vali-

date the value of our new dataset.

2 Related Work

Moral Foundation Theory is a theoretical framework developed by sociologists (Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham et al., 2009) which categorizes moral values into five primary dimensions. It has been extensively employed to examine the influence of moral principles on human behaviors, judgments, decision-making, and political stances (Voelkel et al., 2022; Hoover et al., 2018; Lei and Huang, 2023a; Brady et al., 2017; Lei and Cao, 2023; Fulgoni et al., 2016).

Moral Foundations Lexicon (Frimer, 2019) was widely employed in many applicational studies (Ramezani et al., 2021; Field et al., 2018; Garten et al., 2016) for identifying moral foundations. However, this lexicon match approach suffers from both over-labeling commonly used indicator words and overlooking context-dependent expressions of moral foundations. Researchers (Ramezani et al., 2021) have considered leveraging distributional word embeddings to mitigate the drawbacks of using the Lexicon only.

Annotating Moral Frames for Social Media moral frame annotation efforts have mainly been devoted to social media short text analysis and moral frames have been annotated on Twitter microblogs (Johnson and Goldwasser, 2019; Hoover et al., 2020) and Reddit posts (Trager et al., 2022). Moral frames were usually annotated for an entire social media posts, Roy et al. (2021) further annotates moral roles (individuals or collective enties) for a moral frame. Shahid et al. (2020) is the only prior work we are aware of that annotates moral foundations for news articles, but their annotations were conducted at the sentence level.

Event Subjectivities: a body of prior research have dedicated to studying the subjective aspects of meanings for events. Ding and Riloff (2018); Zhuang et al. (2020); Zhuang and Riloff (2023) acquired subjective events from texts and recognizing their polarities (*positive* or *negative*). Rashkin et al. (2016); Feng et al. (2013) acquired connotation lexicons and connotation frames for event predicates. Deng and Wiebe (2015) annotates and studies events as opinion targets in the context of a news article.

3 EMONA Dataset Construction

3.1 Data Sources

We intentionally select three different components to form our dataset. Overall, the dataset comprises 400 news articles, 10k sentences and 283k words.

- *AllSides* (Baly et al., 2020) provides article-level ideology stance labels: *left*, *center*, and *right*. We select 12 frequently discussed topics: *abortion*, *coronavirus*, *elections*, *gun control*, *immigration*, *lgbt rights*, *race and racism*, *violence*, *politics*, *us house*, *us senate*, *white house*. Each topic comprises five articles for each of the three stances, spanning from 2012 to 2020, resulting in a total of 180 articles.

- *BASIL* (Fan et al., 2019) provides sentence-level media bias labels and collects 100 triples of articles from three medias (Fox News, New York Times, Huffington Post) discussing the same event. We sample five triples for each year from 2010 to 2019, leading to 50 articles from each media and a total of 150 articles.

- *MPQA 3.0* (Deng and Wiebe, 2015) annotates general opinions towards entities and events for 70 articles from the years 2001 to 2002. We retained all 70 articles for the study of the correlation between event-level moral opinions and opinionated events.

3.2 The Annotation Procedure

The annotators were instructed to annotate a plain article in two passes, identifying event mentions and assigning moral labels for events. Before any annotation takes place, we ask our annotators to first read the entire article and understand the author's overall judgement, stance and opinions. Then in the first pass of annotations, our annotators identify all the event mentions in a news article and annotate individual words as event mentions, following the annotation guidelines of Richer Event Description (O'Gorman et al., 2016) in identifying minimum spans of event mentions. In the second pass of annotations, our annotators examine each event mention with respect to its local and global contexts, and identify if the event bears any moral judgement from the author or the source for events appearing in quotations. For the event mentions that carry moral judgements, the annotators will assign one moral dimension to each event mention,

	Pairwise			Majority			llama-2	gpt-3.5	gpt-4
	Min	Max	Mean	Min	Max	Mean	Mean	Mean	Mean
Event Extraction	0.7670	0.8182	**0.7938**	0.8656	0.8877	0.8772	0.4464	0.5870	0.7113
Moral Identification	0.4797	0.5846	**0.5276**	0.6483	0.7814	0.7017	0.3374	0.4064	0.4707
Ten Moral Classification	0.6682	0.7876	**0.7230**	0.8279	0.9245	0.8679	0.3685	0.4068	0.5546

Table 1: Pairwise Cohen's kappa inter-agreement scores among five annotators, and their agreement with majority voting label. The right three columns show the average agreement scores between each large language model (llama-2-7b-chat, gpt-3.5-turbo, and gpt-4) and the five human annotators.

	# Doc	# Sent	# Event	# (%) Moral
ALL	400	10912	45199	9613 (21.27)
AllSides	180	5448	22527	5628 (24.98)
BASIL	150	3811	16200	2586 (15.96)
MPQA 3.0	70	1653	6472	1399 (21.62)

Table 2: Statistics of EMONA and its three portions. % Moral represents the ratio of moral related events in all the annotated events.

the primary dimension out of the ten moral dimensions. Even though we do not explicitly annotate event arguments [2], we ask our annotators to identify the agent and patient of each event mention and consider the intention of the agent and the effect on the patient when determining moral judgements toward an event.

We recruited five annotators[3] and conducted annotation training until a satisfactory level of inter-annotator agreement was reached. Identifying event mentions (the first pass) is relatively straightforward, but recognizing and classifying moral opinions in context (the second pass) has been approved to be difficult and takes rounds of discussions to reach a stable level of agreement among our multiple annotators. In the official annotation process, we first asked all of our annotators to annotate a common set of 25 documents, and then evenly distributed the remaining documents among our annotators and had each annotator label another 75 documents.

3.3 Inter-annotator Agreements (IAAs)

Based on the common set of 25 documents that contains 3,194 events, we calculate Cohen's kappa inter-annotator agreements (IAAs) between each pair of annotators. In Table 1 (the first column), we report the minimum, maximum, and mean of the pairwise agreement scores among five annotators for each of the three tasks, event identification (identifying each word as event or not event), moral identification (identifying an event as moral or non-moral), and ten moral classification (classifying among the ten moral dimensions for events that have been identified as moral). We also report agreements between each annotator and the majority voting labels (the second column). We can see that identifying events is a relatively easy task, and identifying whether an event bears any moral judgment turns out to be more difficult than discerning among the ten moral dimensions.

3.4 Large Language Models vs. Humans

We further investigate the capabilities of large language models (OpenAI, 2023; Touvron et al., 2023) in event morality annotation. The large language models gpt-4, gpt-3.5-turbo, and llama-2-7b-chat were prompted to automatically identify events and assign event-level moral labels for the same set of articles (the used prompt is in Appendix A). The agreement scores between model generated labels and human annotators are presented in the right three columns of Table 1. Compared to the other two large language models, gpt-4 aligns better with human annotators on event identification and event morality identification and classification. But the model vs. human agreement levels are overall still lower than human vs. human agreements, especially in classifying among the ten moral dimensions.

3.5 Dataset Statistics

Table 2 shows the basic statistics of the EMONA dataset and its three portions. In total, 45,199 event mentions were annotated in EMONA and roughly 21 percents of them (9,613 event mentions) were assigned with a moral label. Table 3 further shows the distributions of the ten moral labels. We can see that *Care/Harm*, *Fairness/Cheating* and *Authority/Subversion* are much more frequent than the other two dimensions *Loyalty/Betrayal* and *Purity/Degradation*. Within each of the three most

[2]Annotating event arguments itself is a strenuous task and event extraction is not the main focus of creating this dataset.

[3]Four graduate students and one undergraduate student conducting research in natural language processing

	Care	Harm	Fairnss	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
ALL	518 (5.39)	1815 (18.88)	688 (7.16)	2445 (25.43)	99 (1.03)	97 (1.01)	1252 (13.02)	2510 (26.11)	90 (0.94)	99 (1.03)
AllSides	345 (6.13)	1367 (24.29)	414 (7.36)	1435 (25.50)	50 (0.89)	60 (1.07)	496 (8.81)	1384 (24.59)	50 (0.89)	27 (0.48)
BASIL	104 (4.02)	271 (10.48)	187 (7.23)	738 (28.54)	42 (1.62)	32 (1.24)	471 (18.21)	678 (26.22)	25 (0.97)	38 (1.47)
MPQA 3.0	69 (4.93)	177 (12.65)	87 (6.22)	272 (19.44)	7 (0.50)	5 (0.36)	285 (20.37)	448 (32.02)	15 (1.07)	34 (2.43)

Table 3: Number (Ratio) of events with ten moral labels in the EMONA dataset and its three portions. Blue moral labels represent positive moral opinions, and red moral labels represent negative moral opinions.

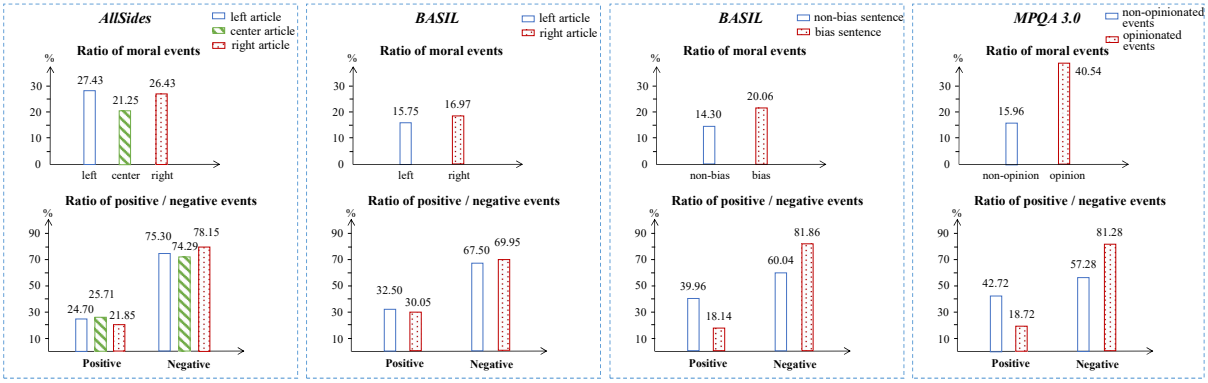

Figure 3: The ratio of moral events in all the annotated events, as well as the ratio of positive / negative events in moral events in different portions of EMONA dataset.

common moral dimensions, the negative class is two or three times more frequent than the positive class. The imbalanced distribution of the ten moral classes poses a challenge for automatic event morality analysis.

4 Dataset Analysis

This section provides a numerical and visual analysis (Figure 3, 7) of the correlations between event-level moral opinions and article-level ideology, sentence-level media bias, and event-level general opinions.

4.1 Analysis with Article-level Ideology

We aim to address the research question of whether event-level moral opinions can reflect article-level ideology. The ratios of moral events across articles from left, center, and right media outlets within the AllSides portion are plotted in the first subfigure of Figure 3. We observe that the articles with center stance have a lower ratio of moral events than the articles with polarized stances. The left or right articles exhibit a higher ratio of negative moral events, while center articles carry more positive moral judgments. In the second subfigure of Figure 3, we also plot the ratio of moral events in the left leaning and right leaning articles within the BASIL portion, where left articles are from Huffington Post and New York Times while right articles are from Fox News. The detailed ratios of moral events distributed among ten moral foundation categories can be found in Appendix B. We can see that the articles with different ideologies showcase different preferences of moral foundations. The analysis validates that our event-level moral opinions annotations can inform article-level ideological bias.

4.2 Analysis with Sentence-level Media Bias

We also explore the research question of whether event-level moral opinions can designate sentence-level political bias. The ratios of moral events across non-bias and bias sentences in the BASIL portion are plotted in the third subfigure of Figure 3. It is evident that bias sentences exhibit a higher ratio of moral events than non-bias sentences. Moreover, bias sentences tend to include more negative moral judgments, indicating that the journalists often influence public opinions through moral criticism. In contrast, non-bias sentences display negative criticism and positive praise more evenly. The analysis confirms that moral opinions towards events can designate and explain sentence-level media bias.

4.3 Analysis with Event-level Opinions

The relation between event-level moral opinions and event-level general opinions is another research question. The fourth subfigure of Figure 3 presents the ratios of moral events among the opinionated

	event moral identification			event moral classification			end-to-end system		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
lexicon	36.49	27.80	31.56	18.21	19.39	16.75	-	-	-
gpt-3.5-turbo	41.45	58.90	48.66	24.20	27.46	22.08	22.98	23.14	20.61
gpt-4	59.25	60.91	60.06	35.08	32.68	30.83	30.64	32.73	30.06
longformer	61.81	65.48	63.59	46.32	36.56	39.50	44.30	35.47	38.42

Table 4: Performance of intrinsic evaluations on EMONA dataset. The last row represents the results of our built evaluation model. Ten-folder cross validation is conducted.

/ non-opinionated events annotated in the MPQA 3.0 dataset. The opinionated events indeed carry a higher ratio of moral judgments, especially negative criticism. The analysis proves the positive correlation between moral judgments and general opinions towards events.

In addition, we also look into the difference between moral events and opinionated events, and show the confusion matrix in Table 5. There are 886 (59.46%) opinionated events not relevant to moral judgments, which implies that not all opinions are moral opinions. For example, *the medical policy hurts the economy*, where the author conveys negative sentiment towards *hurt* event but the judgment is evaluative and not from the moral perspective. Also, we observe that 795 (56.82%) moral events are not identified as opinionated events. This indicates that event-level moral opinions can be more implicit than opinionated content, thereby can supplement implicit opinions. Hence, our event-level moral opinions annotations can uncover and supplement general opinions.

	moral events	non-moral events
opinionated events	604	886
non-opinionated events	795	4147

Table 5: Confusion matrix of the number of moral events and opinionated events in MPQA 3.0 dataset.

5 Intrinsic Evaluation

In this section, we conduct intrinsic evaluations for the newly created dataset EMONA. Specifically, we propose the following tasks and build evaluation models for them: (1) event moral identification: identify whether an event contains moral judgment or not (2) event moral classification: classify moral labels for events, including ten moral labels and *non-moral* label (3) end-to-end system: predict label for every words in a plain article, including ten moral labels, *non-moral* label, and *non-event* label. The former two tasks target towards events, while the end-to-end system generates labels for every words. The intrinsic evaluation models we built can serve as baselines for future work.

5.1 The Baseline Models

Considering the news articles are usually long, we utilize the Longformer (Beltagy et al., 2020) as the language model to encode the entire article. We also add an extra layer of Bi-LSTM (Huang et al., 2015) on top to capture the contextual information. Given an article consisting of n words, the derived word embeddings are $(w_1, \ldots, w_n)$, among which the events embeddings are $(e_1, \ldots, e_m)$.

The event moral identification model builds a two-layer binary classification head on top of the event embedding e_i to make predictions. The event moral classification model also builds an 11-class classification head on top of the event embeddings e_i, to predict the probability of each moral label. The end-to-end system builds a 12-class classification head on top of each word embedding w_i and generates labels for every word. The classical cross-entropy loss is employed for training.

5.2 Experimental Settings

Ten-fold cross validation is performed for evaluation. The entire dataset EMONA is split into ten folders of the equal size. In each iteration, a fold is used as the test set, eight folds are used as the training set while a remaining fold is used as the validation set to determine when to stop training. The evaluation metrics are calculated based on all the ten test folders. Precision, Recall, and F1 score of the positive class are reported for event moral identification task. Macro Precision, Recall, and F1 score are reported for event moral classification and the end-to-end system.

The AdamW (Loshchilov and Hutter, 2019) is used as the optimizer. The maximum training epochs is 10. The learning rate is initialized as 1e-5 and adjusted by a linear scheduler. The weight decay is set to 1e-2.

Error Analysis
Example 1: In a succinct speech, the new president told Americans: "The time for empty **talk [predicted: non-moral; correct: cheating]** is over. Now arrives the hour of **action [predicted: non-moral; correct: authority]**."
Example 2: **Violence [predicted: harm; correct: harm]** flared anew on Saturday night when 100 protesters **attacked [predicted: harm; correct: subversion]** the police with riot gear.

Figure 4: Error analysis of event-level moral opinions classification

5.3 Experimental Results

Table 4 presents the performance of intrinsic evaluations. The last row shows our evaluation models based on longformer. We also implement three systems for comparison: a lexicon matching system, where we match event words with the moral foundations lexicon provided by Frimer (2019), and two large language models gpt-3.5-turbo / gpt-4.

The evaluation models based on longformer perform better than the lexicon matching baseline. It is because event-level moral opinions are context-dependent. The same event mentions in different context can take different moral judgments. The simple word-based matching without considering the context cannot well extract these moral opinions. This suggests future directions of encoding broader contextualized information into the model. Also, the two large language models do not surpass the evaluation model based on fine tuning.

The current performance of the end-to-end system is relatively low, and the primary obstacle is the comprehension of event-level moral opinions. The difficulty of moral classification lies in the imbalanced distribution of ten moral classes, which suggests future directions of improving performance on infrequent classes that lack training data.

5.4 Error Analysis

In addition, we also perform an error analysis for event-level moral opinions classification. Firstly, we observe that failing to recognize implicit moral opinion is one type of error. Take example 1 in Figure 4 as an instance, the author implicitly criticizes *empty talk* and subtly praises *action* event. While the model fails to discover the implicitly conveyed opinions and wrongly selects *non-moral* label. This suggests the future directions of incorporating contextual knowledge and enhancing the capability to extract implicit opinions.

Secondly, failing to select the correct moral dimension is another type of error. Figure 5 shows the confusion matrix of ten moral labels. We can see that the confusions between *Harm*, *Cheating*, and *Subversion* are the primary errors. Take example 2 in Figure 4 as an illustration, the model wrongly predicts *attacked* event as *harm*, by focusing on the semantic of event trigger word, while neglecting the event argument *police* that represents *authority*. This indicates the necessity to encode the model with contextualized semantics.

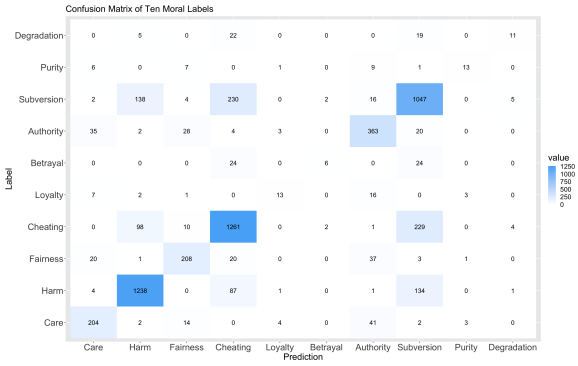

Figure 5: Confusion matrix of ten moral labels

6 Extrinsic Evaluation

To validate the potential applications of this dataset, we further conduct extrinsic evaluations. Motivated by the analysis that confirms positive correlations between moral opinions of events and ideological bias or general opinions, we propose to incorporate event-level moral opinions into three extrinsic evaluation tasks: (1) article-level ideology classification (2) sentence-level media bias identification (3) event-level opinion identification.

6.1 Knowledge Distillation

In particular, we design a knowledge distillation framework to distill the event-level moral knowledge into these downstream tasks, as illustrated in Figure 6. This knowledge distillation framework is able to capture moral opinions for articles with or without event morality golden annotations, and integrate the moral knowledge for various downstream extrinsic evaluation tasks.

Specifically, given a new article that does not have moral annotations, the end-to-end event moral

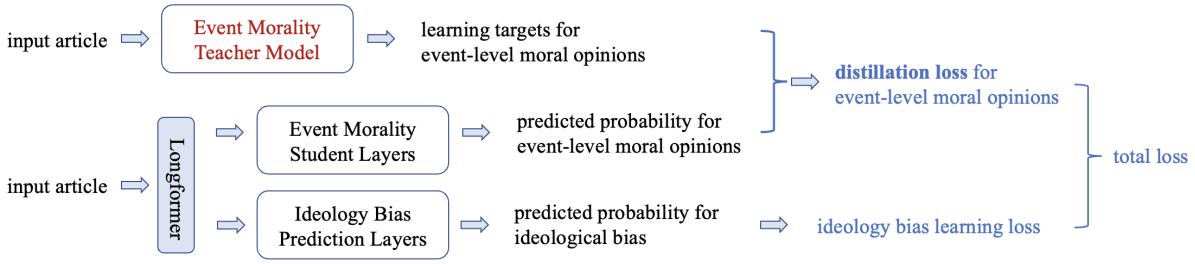

Figure 6: An illustration of ideology bias task informed by event-level moral opinions via knowledge distillation

	article-level ideology			sentence-level media bias			event-level opinions		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
longformer baseline	83.69	84.65	84.11	46.81	45.65	46.22	56.73	57.71	57.21
+ event-level moral opinions	87.15	88.03	87.46	50.79	49.60	50.19	60.93	62.94	61.92

Table 6: Performance of article-level ideology classification on AllSides, sentence-level media bias identification on BASIL, and event-level opinions identification on MPQA 3.0 datasets, informed by event-level moral opinions.

classification system is leveraged as the *teacher model* T_{moral}, and generates the predicted probability of moral labels for each word w_i:

$$P_i = (p_i^1, p_i^2, ..., p_i^{12}) \quad (1)$$

where the predicted probabilities P_i contains the moral knowledge from teacher model T_{moral}, and captures the moral opinions for this new article.

The *student layer* of event morality learning is also built on top of word embeddings w_i, so as to learn the moral knowledge from the teacher model:

$$Q_i = softmax(W_2(W_1 w_i + b_1) + b_2) \quad (2)$$

where the predicted probabilities Q_i is the learned outcome of student layer.

The *distillation loss* is designed to penalize the distance between the learning targets P_i generated by the teacher model and the learned outcome Q_i of the student layer:

$$Loss_{moral} = KL(P, Q) = \sum_{i=1}^{n} P_i log(\frac{P_i}{Q_i}) \quad (3)$$

where KL means the Kullback-Leibler (KL) divergence loss. By minimizing this distillation loss, the moral knowledge from the teacher can be distilled into the model, and the student layers are forced to be updated with event moral knowledge.

The *task-specific layers* are built on top of the encoder. The article-level ideology prediction layers are built on top of the article start token embedding, to predict article ideology into left, center, and right three classes. The sentence-level media bias prediction layers are built on top of the sentence start token embedding, to predict whether the sentence contains political bias or not. The event-level opinions identification layers are built on top of the event embedding e_i, to predict whether the event carries general opinions or not. The classical cross-entropy loss is used to calculate the task-specific loss. The learning objective is the summation of the task-specific loss and the distillation loss.

6.2 Evaluation Datasets

The article-level ideology classification is evaluated on AllSides dataset (Baly et al., 2020). We follow the same splitting setting released by the dataset: 27978 train, 6996 valid, and 1300 test articles. The sentence-level media bias identification is evaluated on the entire 300 articles of BASIL dataset (Fan et al., 2019). We follow the previous work (Lei et al., 2022; Fan et al., 2019) and conduct ten-folder cross validation for evaluation. The event-level opinions identification is experimented on 70 articles in MPQA 3.0 dataset (Deng and Wiebe, 2015), and we also perform ten-folder cross validation due to its small size.

6.3 Experimental Results

The performance of the three extrinsic evaluations are summarized in Table 6. Macro Precision, Recall, and F1 score are reported for article-level ideology classification task. Precision, Recall, and F1 score of the positive class are reported for the other two binary classification tasks.

Incorporating event-level moral opinions can noticeably improve both precision and recall on the three extrinsic tasks, leading to F1 score increased by 3.35% to 4.71%. The probabilistic features

provided in the teacher model contains fuzzy and nuanced moral knowledge, thus can benefit the three downstream tasks. The performance gains demonstrate the usefulness of event-level moral opinions, and validates the potential applications of our newly created dataset EMONA.

7 Conclusion

This paper defines a new task that understands moral opinions towards events in news articles. We have created a new dataset EMONA, and provide the first annotations of event-level moral opinions in news articles. The dataset analysis, intrinsic evaluation, and extrinsic evaluation on three downstream applications are conducted for this new dataset, which showcase that event-level moral opinions can effectively reflect article-level ideology, designate sentence-level political bias, and uncover event-level implicit opinions. For future work, we will continue to improve the performance of event level moral opinion identification and classification, in addition, we are interested to explore other applications of this task, such as stance detection and news summarization.

Limitations

We built the intrinsic evaluation models for event moral identification and classification, along with an end-to-end system. The future work is supposed to explore more sophisticated methodology to identify and classify event-level moral opinions, and incorporate broader contextual information into event morality understanding. Additionally, we propose a knowledge distillation framework to learn ideological bias tasks and event morality task together. Future work necessities the development of more sophisticated methods to incorporate event-level moral opinions with other correlated tasks.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the award IIS-2127746. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Avnika B Amin, Robert A Bednarczyk, Cara E Ray, Kala J Melchiori, Jesse Graham, Jeffrey R Huntsinger, and Saad B Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Selim Berker. 2019. The explanatory ambitions of moral principles. *Noûs*, 53(4):904–936.

William J Brady, Molly J Crockett, and Jay J Van Bavel. 2020. The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4):978–1010.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145(3):366.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

Haibo Ding and Ellen Riloff. 2018. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, New Orleans, Louisiana. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria. Association for Computational Linguistics.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Jeremy A Frimer. 2019. Moral foundations dictionary 2.0.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

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

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

Joe Hoover, Kate Johnson, Reihane Boghrati, Jesse Graham, Morteza Dehghani, and M. Brent Donnellan. 2018. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, 4(1).

Joe Hoover, Guadalupe Portillo-Wightman, Luc Yeh, Soheil Havaldar, Amir M. Davani, Yuening Lin, Benjamin Kennedy, Mohammad Atari, Zeina Kamel, Michael Mendlen, Giovanni Moreno, Cindy Park, Tom E. Chang, Justin Chin, Caitlin Leong, Jennifer Y. Leung, Ani Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

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

Kristen Johnson and Dan Goldwasser. 2019. Modeling behavioral aspects of social media discourse for moral classification. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 100–109, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanyuan Lei and Houwei Cao. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969.

Yuanyuan Lei and Ruihong Huang. 2022. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023a. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023b. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822, Singapore. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Artur Nilsson, Arvid Erlandsson, and Daniel Västfjäll. 2020. Moral foundations theory and the psychology of charitable giving. *European Journal of Personality*, 34(3):431–447.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on*

Computing News Storylines (CNS 2016), pages 47–56, Austin, Texas. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Aida Ramezani, Zining Zhu, Frank Rudzicz, and Yang Xu. 2021. An unsupervised framework for tracing textual sources of moral change. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1215–1228, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Jan G Voelkel, Mashail Malik, Chrystal Redekopp, and Robb Willer. 2022. Changing americans' attitudes about immigration: Using moral framing to bolster factual arguments. *The ANNALS of the American Academy of Political and Social Science*, 700(1):73–85.

Christopher Wolsko. 2017. Expanding the range of environmental values: Political orientation, moral foundations, and the common ingroup. *Journal of Environmental Psychology*, 51:284–294.

Christopher Wolsko, Hector Ariceaga, and Jesse Seiden. 2016. Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65:7–19.

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4654–4663, Hong Kong, China. Association for Computational Linguistics.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1418–1428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2020. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5608–5617, Online. Association for Computational Linguistics.

Yuan Zhuang and Ellen Riloff. 2023. Eliciting affective events from language models by multiple view co-prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3189–3201, Toronto, Canada. Association for Computational Linguistics.

A Prompt for Large Language Models

The designed prompt to provide large language models with annotation examples is: "An event word is the word describing a thing that happens, such as occurrence, action, process, or event state. Please extract the event words from the sentence, only one word for one event. Please also provide one of the moral labels based on the Moral Foundation Theory (non-moral, care, harm, fairness, cheating, authority, subversion, loyalty, betrayal, sanctity, degradation). If the moral sense is vague, choose 'non-moral'. Please mimic the following examples style. Example: "More than 200 people crowded in the forum on Friday." Event words: "crowded (non-moral)". Example: "We show empathy for other people who might choose abortion." Event words: "empathy (care), choose (non-moral), abortion (non-moral)". Example: "Mayor asked New Yorkers to report after the execution-style killing of officers." Event words: "asked (non-moral), report (non-moral), killing (harm)". Example: "Colorado law prohibits discrimination on the basis of sexual orientation." Event words: "prohibits (fairness), discrimination (cheating)". Example: "People participating in the worship comply with the social distancing orders issued by the government." Event words: "worship (non-moral), comply (authority), issued (non-moral)". Example: "FBI director praised the massive manhunt as an extraordinary effort by law enforcement." Event words: "praised (non-moral), manhunt (non-moral), effort (authority)". Example: "The mobs are overthrowing the government." Event words: "overthrowing (subversion)". Example: "It is the Democrats fault for being weak and ineffective with border security and crime." Event words: "weak (subversion), ineffective (subversion)". Example: "They have fealty and allegiance to out country." Event words: "fealty (loyalty), allegiance (loyalty)". Example: "Trump is accused of colluding with Russia." Event words: "accused (non-moral), colluding (betrayal)". Example: "He calls for putting faith at the center." Event words: "calls (non-moral), faith (sanctity)". Example: "I think anyone who would suggest the military mission is not a success does disservice to the sacrifice of Chief Ryan Owens." Event words: "suggest (non-moral), success (non-moral), disservice (degradation), sacrifice (loyalty)". Sentence: "xxx" Event words:",

B Distributions Over Ten Moral Foundations

Figure 7 presents the ratio of moral events distributed within ten moral foundations in different portions of EMONA dataset. The articles with left, center, or right ideology in AllSides exhibit a different preference towards model values. The bias and non-bias sentences in BASIL showcase a different distribution of ten moral classes. The opinionated and non-opinionated events in MPQA 3.0 also present a different moral values.

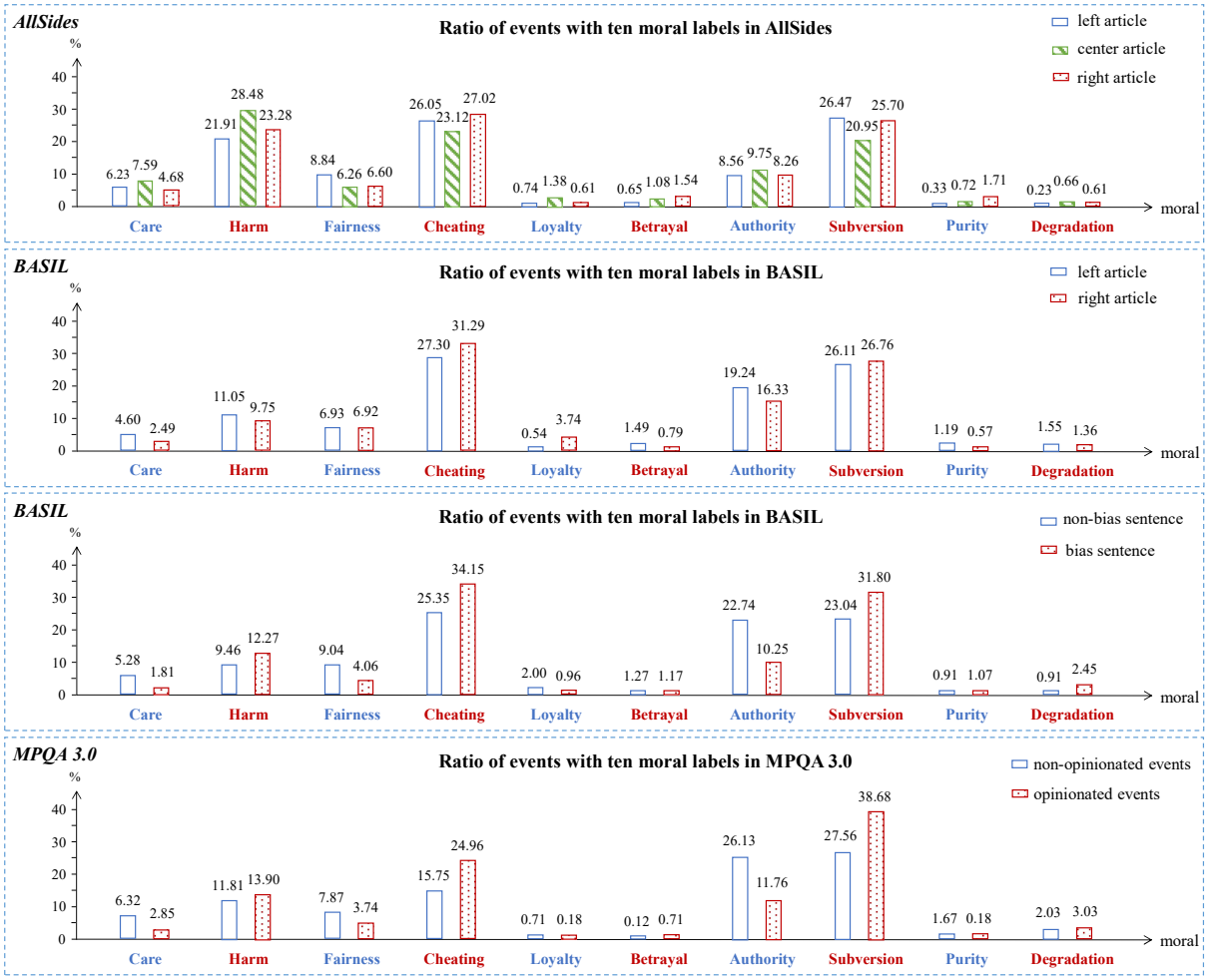

Figure 7: The ratio of moral events distributed within ten moral foundations in different portions of EMONA dataset.