Why Is It Hate Speech?
Masked Rationale Prediction for Explainable Hate Speech Detection

Jiyun Kim, Byounghan Lee, Kyung-Ah Sohn*
Ajou University
{hamjee66, qudgks96, kasohn}@ajou.ac.kr
*Corresponding author

Abstract

In a hate speech detection model, we should consider two critical aspects in addition to detection performance–bias and explainability. Hate speech cannot be identified based solely on the presence of specific words; the model should be able to reason like humans and be explainable. To improve the performance concerning the two aspects, we propose Masked Rationale Prediction (MRP) as an intermediate task. MRP is a task to predict the masked human rationales–snippets of a sentence that are grounds for human judgment–by referring to surrounding tokens combined with their unmasked rationales. As the model learns its reasoning ability based on rationales by MRP, it performs hate speech detection robustly in terms of bias and explainability. The proposed method generally achieves state-of-the-art performance in various metrics, demonstrating its effectiveness for hate speech detection. *Warning: This paper contains samples that may be upsetting.*

1 Introduction

With the recent development of social media and online communities, hate speech, one of the critical social problems, can spread easily. The spread of hate strengthens discrimination and prejudice against the target social groups and can violate their human rights. Moreover, online hatred extends offline and causes real-world crimes. Therefore, properly regulating online hate speech is important to address many social problems related to aversion.

In addition to the detection performance, two essential considerations are involved in implementing a hate speech detection model–*bias* and *explainability*. Hate speech should not be judged by any specific word but by the context in which the word is used. Even if any word generally considered vicious does not exist in a text, the text can be hate speech. A specific expression does not always imply hatred either (*e.g.*, e.g., 'nigger') (Del Vigna12 et al., 2017). However, the presence of this word can cause a model to make a biased detection of hate speech. This erroneous judgment may inadvertently strengthen the discrimination against the target group of the expression (Sap et al., 2019; Davidson et al., 2019). In this respect, the model's *bias* toward specific expressions should be excluded.

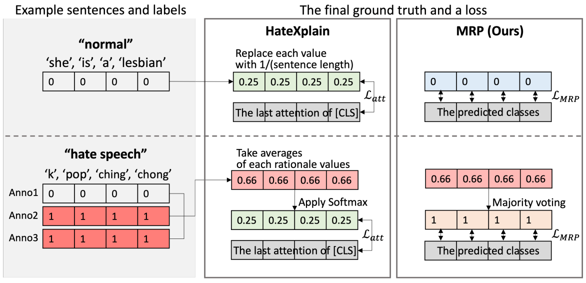


Figure 1: Examples for the two methods to get the final ground truths. Example input sentences are represented with the class and human rationale labels. In this figure, HateXplain uses the same ground truth about both normal and hateful sentences for the loss. However, our method could determine the two classes with the ground truths.

The expressions that can cause biased judgment should be interpreted in context. It means it is vital for the hate speech detection models to have the ability to make judgments based on context, as humans do. Therefore, the model should be *explainable* to humans so that the rationale behind a result is explained (Liu et al., 2018). Here, the rationale is a piece of a sentence as justification for the model's prediction about the sentence, as defined by related research (Hancock et al., 2018; Lei et al., 2016).

To the best of our knowledge, HateXplain (Mathew et al., 2020) is the first hate speech detection benchmark dataset that considers both these aspects. They proposed a method that utilizes rationales as attention ground truths to complement the performance of the two elements. However, when most tokens are annotated as the human rationale in a hateful sentence, the rationale's information

Proceedings of the 29th International Conference on Computational Linguistics, pages 6644–6655
October 12–17, 2022.

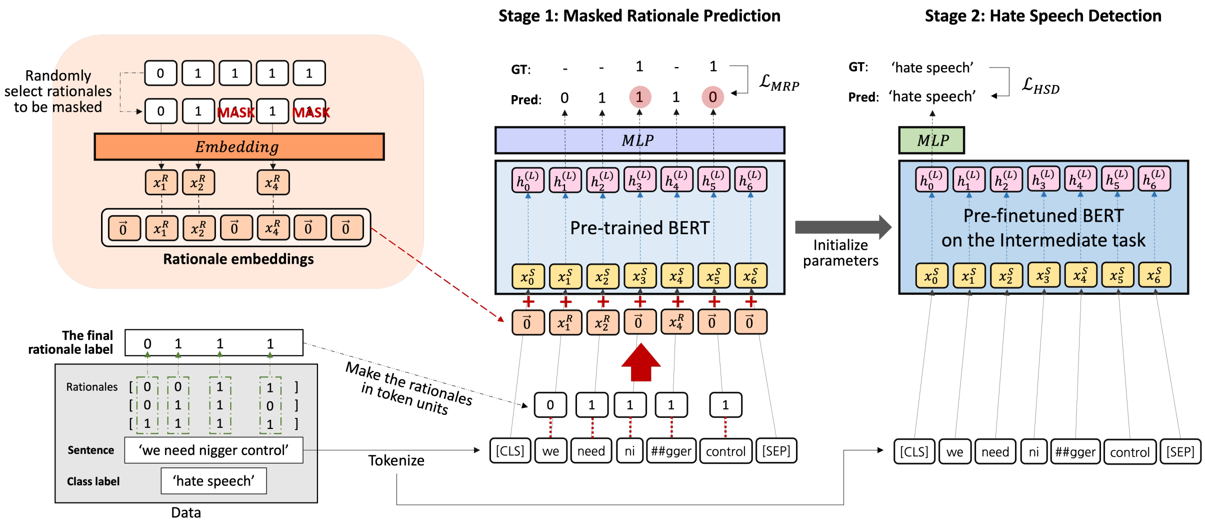


Figure 2: Framework of the proposed method. We finetune a pre-trained BERT through two training stages–Masked Rationale Prediction (MRP) and then hate speech detection. In MRP, the partially masked rationale label is inputted as the rationale embeddings by being added into the input embeddings of BERT. The model predicts each masked rationale per token. The model for hate speech detection is initialized by the updated parameters during MRP.

could be meaningless as the ground truth attention becomes hard to be distinguished from that of a normal sentence, as shown in Figure 1. This can hinder the model's learning.

In this paper, we present a method to implement a hate speech detection model much more effectively by using the human rationale of hate for finetuning a pre-trained language model. To achieve this, we propose Masked Rationale Prediction (MRP) as an intermediate task before finetuning the model on hate speech detection. MRP trains a model to predict the human rationale label of each token by referring to the context of the input sentence. The model takes the human rationale information of some input tokens among the sentence along with the corresponding tokens as input. It then predicts the rationale of the remaining tokens on which the rationale is masked. We embed the rationales to provide the human rationales as input per token. The masking process of the partial rationales is implemented while creating rationale embeddings; some rationale embedding vectors are replaced with zero vectors.

MRP allows the model to make judgments per token about its masked rationale by considering surrounding tokens with an unmasked rationale. With this, the model learns a human-like reasoning process to get context-dependent abusiveness of tokens. The model parameters trained on MRP become the initial parameter values for hate speech detection in the following training stage. In this way, based on the way of human reasoning for hate, the model can get improved abilities in terms of bias and explainability in detecting hate speech. We experimented with BERT (Devlin et al., 2018) as the pre-trained model. Consequently, our models finetuned in the proposed way–BERT-MRP and BERT-RP–achieve state-of-the-art performance overall on all three types of 11 metrics of HateXplain benchmark–Performance-based, Bias-based, and Explainability-based (Mathew et al., 2020). And the two models, especially BERT-MRP, also show the best results in qualitative evaluation of explicit and implicit hate speech detection.

The main contributions of this paper are:

- We propose a method to utilize human rationales as input by transforming them into rationale embeddings. Combining the embedded rationales with the corresponding input sentence can provide information about the human rationales per token during model training.

- We propose Masked Rationale Prediction (MRP), a learning method that leads the model to predict the masked rationale by considering the surrounding tokens. The model is allowed to learn the reasoning process in context.

- We finetune a pre-trained BERT in two stages–on MRP as an intermediate task and then on hate speech detection. The parameters trained concerning human reasoning for hate become a sufficient basis not only for the detection but also in terms of the model bias and explainability.

2 Related works

Hate speech detection With the advance of deep learning, hate speech detection studies have utilized neural networks (Badjatiya et al., 2017; Han and Eisenstein, 2019), and word embedding methods (McKeown and McGregor, 2018). More recently, Transformer-based (Vaswani et al., 2017) models have shown remarkable results. In hate speech detection, BERT has been adopted for various studies as hate speech detection can be considered a classification task. (Mandl et al., 2019; Ranasinghe et al., 2019) compared a BERT-based model with Recurrent Neural Networks (RNNs)-based models and showed the BERT-based model outperforms other models. Furthermore, some studies have considered the model's bias and explainability. (Vaidya et al., 2020) improved accuracy and reduced unintended bias by adopting multi-task learning that predicts toxicity of text and target group labels as additional information. (Mathew et al., 2020) utilized rationales of the dataset as additional information for finetuning BERT to deal with the bias and explainability. To improve performance in terms of the two considerations, we propose a more effective finetuning approach based on BERT and the rationales by adopting the pre-training framework.

Pre-finetuning on an intermediate task Recently, finetuning a pre-trained model on a downstream task has become the norm (Howard and Ruder, 2018; Radford et al., 2018). However, it cannot be guaranteed that the model finetuned with a small dataset compared to its size will be sufficiently well-adjusted for the target downstream task (Phang et al., 2018). Pre-finetuning is a technique to train the model on a task before the target task (Aghajanyan et al., 2021). This can help the model learn the data patterns or reduce the tuning time so that it converges quickly to better fit the target task. According to (Pruksachatkun et al., 2020; Aghajanyan et al., 2021), the more closely the intermediate task is related to the target task, the better the effect of pre-finetuning. And inference tasks involving the reasoning process show a remarkable improvement in the target task performance. We adopt this method to train a pre-trained language model through two stages for hate speech detection. As the intermediate task, we propose MRP, which guides the model to infer the human rationale of each token based on surrounding tokens.

Explainable NLP and rationale Explaining the rationale of the result of an AI model is necessary for it to be explainable to humans (Liu et al., 2018). Some Natural Language Processing (NLP) studies define rationale as snippets of an input text on which the model's prediction is supported (Hancock et al., 2018; Lei et al., 2016). (Lei et al., 2016) implemented a generator that generates words considered rationales and used them as input of an encoder for sentiment classification. (Bao et al., 2018) mapped the human rationales into the model attention values to solve the low-resource problem by learning a domain-invariant token representation. For hate speech detection, HateXplain employs the human rationales as ground truth attention to concentrate on aggressive tokens. Unlike existing approaches, we utilize the masked human rationale label embeddings as input. They become the useful additional information of each token.

Masked label prediction The UniMP model presented by (Shi et al., 2020) aims to solve the graph node classification problem using graph transformer networks (Yun et al., 2019). They maximized the propagation information required to reconstruct a partially observable label by using both feature information and label information as inputs. However, to prevent overfitting due to excessive information, some label information is masked, and the masked label is predicted. We apply a similar method to text data for an intermediate task with rationales. Through additional rationale information, the model increases the understanding of input sentences, and the performance of the downstream task is improved.

3 Method

Hate speech detection can be described as a text classification problem. Following the problem setting of HateXplain, we define the problem as a three-class classification involving three categories–'hate speech,' 'offensive,' or 'normal'. We finetune a pre-trained BERT on hate speech detection. Note that other transformer encoder-based models can be used instead. Before finetuning the model on this task, we pre-finetune it on an intermediate task. We propose Masked Rationale Prediction (MRP) as the intermediate task. Our method is described in Figure 2.

3.1 Masked rationale prediction

For MRP, we utilize human rationales of hate provided by the HateXplain dataset. Annotators marked some words in a sentence as rationales for judging the sentence as abusive. A rationale label is presented in a list format, including 1 as rationale and 0 as non-rationale per word in the corresponding sentence. There are no such labels for a sentence whose final class is 'normal.' As the dataset was annotated by two or three people per sentence, some pre-processing is required to get the final rationale labels for MRP. To manipulate the multiple rationale labels to one per sentence, we obtain the average value of the rationales per word, and if it is over 0.5, the value of 1; otherwise, the value of 0 is determined as the final rationale of the corresponding word. The final rationale label is a list of these last values. In the case of the 'normal' sentence, a list of zeros is used. Accordingly, the final rationale label consists of as many 0s or 1s as the number of words in the sentence. As a sentence is tokenized, its rationale label is also modified in token units.

MRP is based on token classification, which predicts the rationale label R per token in an input sentence S. In our MRP, the rationale labels, as well as the sentences, are used as inputs. The process of embedding S is the same as that of BERT. We denote the embedded S as $X^S = \{x_0^S, x_1^S, \cdots, x_{n-1}^S\} \in \mathbb{R}^{n \times d}$ where n is the sequence length and d is the embedding size. And to use R as input, we pass it through an embedding layer to get $X^R = \{x_0^R, x_1^R, \cdots, x_{n-1}^R\} \in \mathbb{R}^{n \times d}$ as shown in Figure 2. The rationale embeddings reflect the attributes of each token as a ground for the human judgment.

MRP differs from BERT's Masked Language Modeling (MLM) in masking processing. Specifically, we do not mask the tokens; we mask the rationales. To construct the partially masked rationale embeddings $\tilde{X}^R$, some rationales are randomly selected to be masked. Each of rationales is transformed into its corresponding embedding vector, except the masked ones. For masking, zero vectors replace the embedding vectors of each corresponding token. For example, if we mask x_1^R and x_3^R, then the rationale embedding matrix is like $\tilde{X}^R = \{\vec{0}, \vec{0}, x_2^R, \vec{0}, \cdots, x_{n-2}^R, \vec{0}\}$. The first and last rationale embeddings corresponding to CLS and SEP tokens, respectively, are replaced with $\vec{0}$.

The MRP model predicts the rationale by taking the sum of the embedded tokens X^S and the partially masked rationales $\tilde{X}^R$ as input. We then get:

$$\begin{aligned} H_{MRP}^{(0)} &= X^S + \tilde{X}^R, \\ H_{MRP}^{(l+1)} &= \text{Transformer}(H_{MRP}^{(l)}), \\ \hat{X}^R &= \text{MLP}(H_{MRP}^{(L)}). \end{aligned} \tag{1}$$

The l-th hidden state passes through the transformer block to create the $l+1$-th hidden state, and the last hidden state $H_{MRP}^{(L)}$ outputs a predicted rationale $\hat{X}^R$ through Multi-Layer Perceptron (MLP). In other words, the model is guided to predict the masked rationales by referring to the representations of tokens using their corresponding observed rationales.

The loss $\mathcal{L}_{MRP}$ is calculated with only the predictions of the masked rationales. Therefore, our objective function is:

$$\begin{aligned} \arg\max_{\theta}\ \log \mathrm{p}_\theta \left(\hat{X}^R | X^S, \tilde{X}^R\right) = \\ \sum_{m \in M} \log \mathrm{p}_\theta \left(x_m^R | X^S, \tilde{X}^R\right), \end{aligned} \tag{2}$$

where M indicates a set of index numbers of rationales that have been masked.

3.2 Hate speech detection

Hate speech detection is implemented as three-class text classification. The model predicts which category Y the input sentence belongs to among 'hate speech', 'offensive', and 'normal'. The head that outputs the predicted class $\hat{Y}$ is used on the top of BERT. Before training, the model parameters are initialized by parameters updated on the intermediate task MRP, except for the head. As the forms of heads are different for two stages, their parameters are randomly initialized. Consequently, in the finetuning stage on hate speech detection, the rationale labels are not involved functionally, considered as $[0]_{n \times d}$. Therefore, in this stage, the input of the model is $H_{HSD}^{(0)} = X^S$.

In this stage, the model does not refer to the rationale labels. The parameters trained during MRP are utilized as a base for reasoning hatefulness in context. The loss $\mathcal{L}_{HSD}$ is obtained through a cross-entropy function, as the task is a multi-class classification problem.

$$\arg\max_{\theta}\ \log \mathrm{p}_\theta \left(\hat{Y} | X^S\right). \tag{3}$$

Model			**Performance**			**Bias**		
	ration.	pre-fin.	Acc.	Macro F1	AUROC	GMB-Sub.	GMB-BPSN	GMB-BNSP
BERT			69.0	67.4	84.3	76.2	70.9	75.7
BERT-HateXplain	✓		69.8	68.7	85.1	80.7	74.5	76.3
BERT-MLM		✓	70.0	67.5	<u>85.4</u>	79.0	67.7	80.9
BERT-RP	✓	✓	**70.7**	<u>69.3</u>	85.3	<u>81.4</u>	<u>74.6</u>	<u>84.8</u>
BERT-MRP	✓	✓	<u>70.4</u>	**69.9**	**86.2**	**81.5**	**74.8**	**85.4**

Table 1: Results for the performance-based and the bias-based metrics. Scores in bold type are the best for each corresponding metric, while the underlined are the second best, and so are in Table 2.

Model			**Explainability**				
			Plausibility			**Faithfulness**	
	ration.	pre-fin.	IOU F1	Token F1	AUPRC	Comp.	Suff. ↓
BERT [Att]			13.0	49.7	<u>77.8</u>	44.7	5.7
BERT [LIME]			11.8	46.8	74.7	43.6	0.8
BERT-HateXplain [Att]	✓		12.0	41.1	62.6	42.4	16.0
BERT-HateXplain [LIME]	✓		11.2	45.2	72.2	**50.0**	0.4
BERT-MLM [Att]		✓	13.5	43.5	60.8	40.1	11.9
BERT-MLM [LIME]		✓	11.3	47.2	76.5	43.4	**-5.5**
BERT-RP [Att]	✓	✓	<u>13.8</u>	<u>50.3</u>	73.8	45.4	7.2
BERT-RP [LIME]	✓	✓	11.4	49.3	77.7	<u>48.6</u>	<u>-2.6</u>
BERT-MRP [Att]	✓	✓	**14.1**	**50.4**	74.5	47.9	6.7
BERT-MRP [LIME]	✓	✓	12.9	50.1	**79.2**	48.3	-1.2

Table 2: Results for the explainability-based metrics. The lower the score Sufficiency in Faithfulness, the better, and the higher the other scores, the better.

4 Experiments

4.1 Dataset

For both stages of the intermediate and the target task, we use the HateXplain dataset. It contains 20,148 items collected from Twitter and Gab. Every item consists of one English sentence with its own ID and annotations about labels for its category, target groups, and rationales, which are annotated by two or three annotators. Based on the IDs, the dataset is split into 8:1:1 for training, validation, and test phases. Following the permanent split provided by the dataset, the models can't reference any test data during the training phases of all stages.

4.2 Metrics

The evaluation is according to the metrics of HateXplain, which are classified into three types: performance-based, bias-based, and explainability-based. The performance-based metrics measure the detection performance in distinguishing among three classes (i.e., hate speech, offensive, and normal). Accuracy, macro F1 score, and AUROC score are used as the metrics.

The bias-based metrics evaluate how biased the model is for specific expressions or profanities easily assumed to be hateful. HateXplain follows AUC-based metrics developed by (Borkan et al., 2019). The model classifies the data into 'toxic'–hateful and offensive–and 'non-toxic'–normal. For evaluating the model's prediction results, the data are separated into four subsets: D_g^+, D_g^-, D^+, and D^-. The target group labels are considered standard for dividing data into subgroups. The notations with g denote the data of a specific subgroup among the subgroups, and the notations without g are the remaining data. $+$ and $-$ mean that the data are toxic and non-toxic, respectively. Based on these subsets, three AUC metrics are calculated.

Subgroup AUC is to evaluate how biased the model is to the context of each target group: $AUC(D_g^- + D_g^+)$. The higher the score, the less biased the model is with its prediction of a certain social group.

BPSN (Background Positive, Subgroup Negative) AUC measures the model's false-positive rates regarding the target groups: $AUC(D^+ + D_g^-)$. The higher the score is, the less a model is likely to confuse non-toxic sentences whose target is the specific subgroup and toxic sentences whose target is one of the other groups.

BNSP (Background Negative, Subgroup Positive) AUC measures the model's false-negative rates regarding the target groups: $AUC(D^- + D_g^+)$. The higher the score is, the less the model is likely to confuse non-toxic sentences whose target is the specific group and toxic sentences whose target is one of the other groups.

We calculate GMB (Generalized Mean of Bias)[1] of the three metrics as the final scores to combine those ten scores of each of the metrics into one overall measure according to the HateXplain benchmark. The formula is: $M_p(m_s) = (\frac{1}{N}\sum_{s=1}^{N} m_s^p)^{\frac{1}{p}}$, where M_p means the p^{th} power-mean function, m_s is one of the bias metrics m calculated for a specific subgroup s, and N is the number of subgroups which is 10 in this paper.

The explainability-based metrics evaluate how much the model is explainable. HateXplain follows ERASER (DeYoung et al., 2019), which is a benchmark for the evaluation of explainability of an NLP model based on rationales. The metrics are divided into Plausibility and Faithfulness. Plausibility refers to how the model's rationale matches the human rationale. Plausibility can be considered both discrete selection and soft selection. For discrete selection, We convert token scores to binary values by more than some threshold(here 0.5). Then, We measures IOU F1 score and Token F1 score. For soft selection, We constructed AUPRC by sweeping a threshold over token scores.

Faithfulness evaluates the influence of the model rationale on its prediction result and consists of Comprehensiveness and Sufficiency. Comprehensiveness assumes the model prediction is less confidence when rationales are removed. This metric can be calculated: $m(x_i)_j - m(x_i \backslash r_i)_j$. $m(x_i)_j$ is the prediction probability of the corresponding class j with an input sentence x_i by the model m. And $x_i \backslash r_i$ is the sentence manipulated by removing the predicted rationale tokens r_i from x_i.[2] The higher a score, the more influential the model's rationales in its prediction. Sufficiency captures the extent to which extracted rationales are acceptable for a model to make a prediction: $m(x_i)_j - m(r_i)_j$. A low score of this metric means that the rationales are adequate in the prediction.

In addition, for the HateXplain benchmark, the scores are calculated based on the attention scores of the last layer or by using the LIME method (Ribeiro et al., 2016). The former is marked as [Att], and the latter is [LIME] in Table 2. (DeYoung et al., 2019) and (Mathew et al., 2020) contain more detailed explanations.

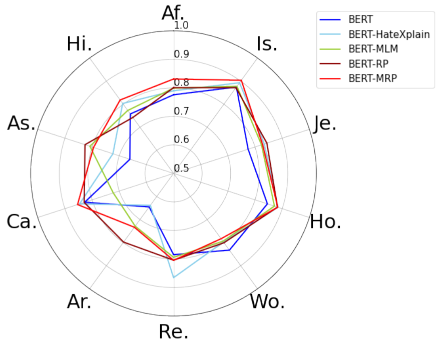


Figure 3: The Subgroup scores among bias-based metrics for each of ten target groups. The target group labels are 'African', 'Islam', 'Jewish', 'Homosexual', 'Women', 'Refugee', 'Arab', 'Caucasian', 'Asian', and 'Hispanic' in clockwise direction respectively. The BPSN and BNSP scores are attached in Appendix.

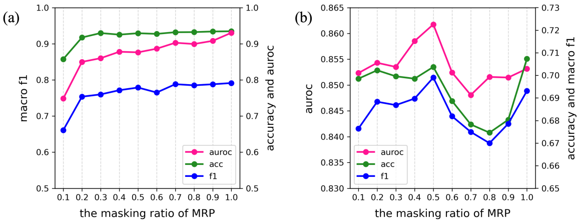


Figure 4: Classification test scores of the proposed models according to masking ratio in MRP. (a) is for token classification after training on MRP in the first stage, and (b) is for hate speech detection in the final stage. The case of masking 100% of tokens is the same as BERT-RP.

4.3 Models and Experimental settings

The evaluated models in Table 1 and Table 2 are as follows. All models are based on a BERT-base-uncased model for a pre-trained model and finetuned on hate speech detection. **BERT** in the tables is simply finetuned on hate speech detection with a fully-connected layer as a head for the three-class classification described above.

BERT-HateXplain uses attention supervision in addition to BERT. It matches the last attention values corresponding to the CLS token to the rationale used as ground truth attention. With this, the CLS token takes additional rationale-based attention information for the prediction. The loss is the summation of this attention loss and the detection loss. The results of BERT and BERT-HateXplain are the same as those presented in (Mathew et al., 2020).

BERT-MLM is evaluated to compare the effectiveness of pre-finetuning tasks. Training a pre-trained NLP model with MLM using data of the

[1] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

[2] We select the top 5 tokens to remove based on the average length of human-annotated rationale labels in the dataset according to HateXplain benchmark.

downstream task is frequently used for the model to understand the downstream data and improve its performance (Han and Eisenstein, 2019; Ben-David et al., 2020; Arefyev et al., 2021). It is implemented by simply masking 15% tokens of each input sentence.

BERT-MRP and **BERT-RP** are the proposed models in this paper. **BERT-MRP** is the model trained on MRP as an intermediate task and then finetuned on hate speech detection. The ratio of masked rationales per token is set to 50% of the entire rationale label. **BERT-RP** is trained on Rationale Prediction (RP), which is MRP when the ratio is set to 100%–masking all the rationales. It is functionally the same as token classification with the rationale label as ground truth.

BERT-MLM, BERT-RP, and BERT-MRP are directly trained in this study. The experimental settings are the same for all models and each training step. The learning rate is $5e{-}5$ during pre-finetuning and $2e{-}5$ for hate speech detection, which is the same as BERT-HateXplain. We use the RAdam optimizer and an Nvidia GeForce GTX 1050 graphics card.

4.4 Comparisons of results

Table 1 and Table 2 present the performances of the models. For all metrics, the proposed models–the two from the bottom–perform much better overall.

Performance-based metrics As summarized in Table 1, the proposed method outperforms the other methods. BERT-MRP shows the highest scores for Macro F1 and AUROC and BERT-RP for accuracy. The pre-finetuned models perform better than those that are not. It shows that the pre-finetuning process helps understand the data and allows enough time for tuning parameters for the target task, thereby improving performance. On the other hand, among the pre-finetuned models, the proposed models show better results than BERT-MLM. Furthermore, they outperform BERT-HateXplain, which also uses the rationale during training like ours. This shows that predicting the human rationale for hate as an intermediate task effectively implements a hate speech detection model.

Bias-based metrics For the model bias, the proposed models show superior results compared to other models. According to Table 1, the models trained using the rationales achieve higher scores than others in general. Given that the human rationales of hatred imply that hate speech is judged based on the context, not merely specific expressions, learning the rationale can exclude the model bias towards the particular words for the prediction. Withal, the proposed BERT-RP and BERT-MRP show better performance than BERT-HateXplain, even though they all utilize the rationale in training. BERT-MRP shows the best scores, and BERT-RP is the second-best for all three metrics. Additionally, Figure 3 shows the scores of the models for each of the ten major target groups. It can be seen that the proposed models score evenly high for all the target groups. While other models have significant differences in their bias depending on the groups, the proposed models have comparatively no correlation with them.

Explainability-based metrics In terms of explainability, the proposed models still perform better than others overall. For Plausibility, BERT-MRP achieves the best performance for all three metrics. It scores much higher than others because it is allowed to directly guess the human rationales during the intermediate training stage. For Faithfulness, BERT-HateXplain[LIME] shows the highest score for Comprehensiveness, and BERT-MLM[LIME] is the best for Sufficiency. However, these models do not reliably score well when considering all four scores obtained according to each of the two measurement methods: attention values or LIME. They show worse scores than BERT for the rest of the scores. On the other hand, BERT-MRP and BERT-RP offer stably high performance for all five explainability-based metrics.

Based on all these results, BERT-MRP and BERT-RP demonstrate the best performance overall for the three types of metrics. Thus, learning the human rationale as an intermediate task before training on hate speech detection seems effective for detection performance and model bias and its explainability. This framework contributes to better performance than the other–pre-finetuning on MLM as well as another way of using rationales.

BERT-MRP generally achieves better results than BERT-RP, wherein the intermediate task is basically the same as token classification. The plots in Figure 4 show the change in test scores according to the masking ratio in MRP. According to Figure 4(a), when more than 20% of all rationales were masked, there is no significant difference in the token classification performance, although the amount of loss decreases as the ratio decreases. When each model was re-trained for hate speech

No.	Model	Example sentence and Rationale	Label
Case 1 Explicit hate speech			
1	Human	imagine unironically believing the high iq kike meme	HS
	BERT-MRP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-RP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-HX	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
2	Human	<user> ffs these ghetto blacks are ignorant like cracker mountain hillbillies	HS
	BERT-MRP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-RP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-HX	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
Case 2 Implicit hate speech			
3	Human	yah christian island republic shall not have queers or witches on it if we find any death	HS
	BERT-MRP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-RP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-HX	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	NO
4	Human	you can not culturally enrich a moslem	HS
	BERT-MRP	you can not culturally en ##rich a mo ##sle ##m	HS
	BERT-RP	you can not culturally en ##rich a mo ##sle ##m	NO
	BERT-HX	you can not culturally en ##rich a mo ##sle ##m	NO
5	Human	<user> he is infected with jihadi virus he will spread it to others	HS
	BERT-MRP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-RP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-HX	<user> he is infected with jihad ##i virus he will spread it to others	OF

Table 3: The highlighted words of the human rationale and the rationale of the models with detection results. BERT-HX is BERT-HateXplain. In the label column, the ground truth is of humans and the remaining labels are the predictions of each model. HS is 'hate speech,' OF is 'offensive,' and NO is 'normal.' More examples are in Appendix.

detection, as shown in Figure 4(b), the case of 50% ratio in BERT-MRP achieved the best classification performance. As MRP is a method for inferring the rationale of a particular token based on surrounding tokens, the model can successfully learn the human rationale within the context. Learning parameters during this reasoning process based on context seems to effectively prevent biased prediction while still being explainable and consequently improves the detection performance substantially.

4.5 Qualitative results

Table 3 shows examples of detection results from models that use human rationale for their training. The visualized values as the model rationales are the LIME results used to measure the explainability-based scores. For the human ground truth, the average value per word of human-annotated rationales is expressed for each word. The darker the color, the higher the values.

It is relatively easy to judge explicit hate speech that includes clear derogatory expressions. As shown in Case 1 of Table 3, all the models perform well. The human rationale tends to focus on specific abusive words, and so does the rationale of each model. However, the rationales of the two proposed models match the ground truth better than BERT-HateXplain. Our method to train a model on a token classification-based task leads well the model to focus on human-like grounds in the sentence by directly learning the human rationale.

As in Case 2, the implicit hate speech with no aggressive expressions cannot be grasped through context. The human rationale thus tends to appear throughout the sentence. As this might make hate speech detection relatively challenging, the detection results of BERT-HateXplain or BERT-RP seem incorrect for some sentences. However, BERT-MRP works accurately based on its rationale that is much more similar to human's than others. Meanwhile, BERT-HateXplain shows a low matching rate of the rationale when the human rationale is throughout the sentence. It uses the human rationale as the ground truth attention, and if there is no difference in the human rationale across tokens, the ground truth could become similar to that of any normal sentence represented by uniform values. This affects the model's explainability and may lead to incorrect detection results. The proposed method does not cause that problem. It gets the distinguishable ground truth from normal ones and assigns it as labels to tokens. On the other hand, the rationale of BERT-MRP matches the ground truth better than that of BERT-RP. MRP requires more context-awareness ability when predicting the masked token by allowing the model to consider the abusiveness of surrounding words that are provided corresponding human rationale. This offers robust detection performance, even when it is necessary to understand the context.

5 Conclusion

This paper presents a method to implement a hate speech detection model considering bias and explainability. We adopt a framework to finetune a pre-trained language model in two stages. As the intermediate task, we propose Masked Rationale Prediction (MRP), which predicts masked rationales for some tokens with the additional rationale information of the remaining surrounding tokens. With this, the model learns to identify abusiveness for each token and the human reasoning process based on context. The trained model by MRP is finetuned again on hate speech detection.

As a result, across quantitative and qualitative evaluations, the proposed model shows state-of-the-art performance in bias and explainability, as well as the detection result. And the examples demonstrate its robustness in detecting hate speech, whether explicit or implicit, based on its superior explainability. Meanwhile, we experimented with only BERT as the pre-trained model to compare our method with base models. But any other transformer encoder-based model can be easily applied, which can be taken as future work.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF) (NRF-2021S1A5A2A03065899), and also by the NRF of Korea grant funded by the Korean government (MSIT) (NRF-2022R1A2C1007434).

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov. 2021. Nb-mlm: Efficient domain adaptation of masked language models for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9114–9124.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. Perl: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv preprint arXiv:1904.02817*.

Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Hui Liu, Qingyu Yin, and William Yang Wang. 2018. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Rohan Kshirsagar1 Tyrus Cukuvac1 Kathleen McKeown and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *EMNLP 2018*, page 26.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (working notes)*, pages 199–207.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

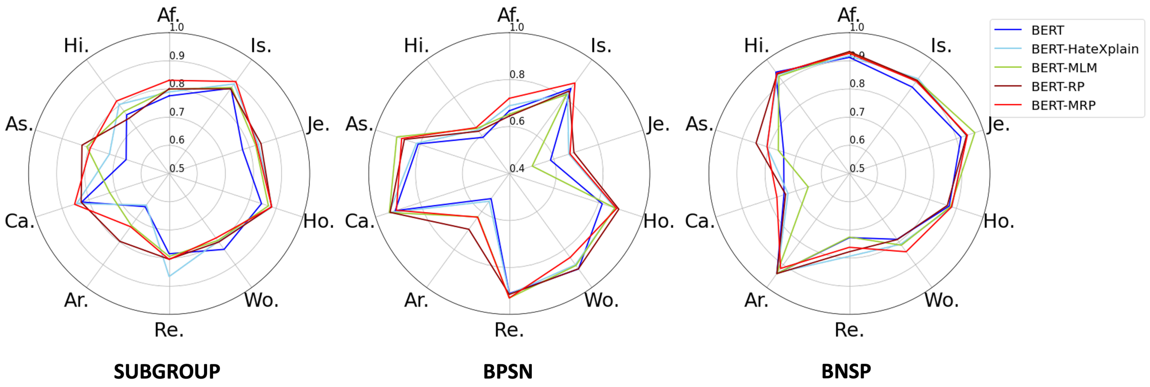
Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and A Noah Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2020. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in neural information processing systems*, 32.



SUBGROUP BPSN BNSP

Figure A1: The Bias-based scores for each 10 target groups. The target group labels are ‘African’, ‘Islam’, ‘Jewish’, ‘Homosexual’, ‘Women’, ‘Refugee’, ‘Arab’, ‘Caucasian’, ‘Asian’, and ‘Hispanic’ in clockwise direction respectively.

Model	**Example sentence and Rationale**	**Label**
Human	if you ’ re an army that is republican you are not welcome here we hate you	HS
BERT-MRP	if you ’ re an army that is republican you are not welcome here we hate you	HS
BERT-RP	if you ’ re an army that is republican you are not welcome here we hate you	NO
BERT-HX	if you ’ re an army that is republican you are not welcome here we hate you	NO
Human	the german regime is more scared of women like this than than are of any moslem terrorist	HS
BERT-MRP	the german regime is more scared of women like this than than are of any mo ##sle ##m terrorist	HS
BERT-RP	the german regime is more scared of women like this than than are of any mo ##sle ##m terrorist	HS
BERT-HX	the german regime is more scared of women like this than than are of any mo ##sle ##m terrorist	NO
Human	i poisoned sergei skripal he was a faggot anyway	HS
BERT-MRP	i poisoned sergei sk ##rip ##al he was a fa ##gg ##ot anyway	HS
BERT-RP	i poisoned sergei sk ##rip ##al he was a fa ##gg ##ot anyway	HS
BERT-HX	i poisoned sergei sk ##rip ##al he was a fa ##gg ##ot anyway	OF
Human	<user> would have pressed that bitch and beat the faggot out of him	HS
BERT-MRP	<user> would have pressed that bitch and beat the fa ##gg ##ot out of him	HS
BERT-RP	<user> would have pressed that bitch and beat the fa ##gg ##ot out of him	HS
BERT-HX	<user> would have pressed that bitch and beat the fa ##gg ##ot out of him	OF
Human	massive muzrat espionage	HS
BERT-MRP	massive mu ##z ##rat espionage	HS
BERT-RP	massive mu ##z ##rat espionage	HS
BERT-HX	massive mu ##z ##rat espionage	HS
Human	enjoy your kike faith	HS
BERT-MRP	enjoy your ki ##ke faith	HS
BERT-RP	enjoy your ki ##ke faith	HS
BERT-HX	enjoy your ki ##ke faith	HS

Table A1: The highlighted words of the human rationale and the rationale of the models with detection results in ’hate speech’ sentences.

Model	Example sentence and Rationale	Label
Human	“ i don ’ t have to respect women i ’ m gay ” kevin	OF
BERT-MRP	“ i don ’ t have to respect women i ’ m gay ” kevin	OF
BERT-RP	“ i don ’ t have to respect women i ’ m gay ” kevin	NO
BERT-HX	“ i don ’ t have to respect women i ’ m gay ” kevin	NO
Human	logan paul is a fucking bozo	OF
BERT-MRP	logan paul is a fucking bo ##zo	OF
BERT-RP	logan paul is a fucking bo ##zo	NO
BERT-HX	logan paul is a fucking bo ##zo	NO
Human	<user> <user> there are literally nazis all over all the time rigjt now throw a rock you hit a nazi	OF
BERT-MRP	<user> <user> there are literally nazis all over all the time rig ##j ##t now throw a rock you hit a nazi	OF
BERT-RP	<user> <user> there are literally nazis all over all the time rig ##j ##t now throw a rock you hit a nazi	OF
BERT-HX	<user> <user> there are literally nazis all over all the time rig ##j ##t now throw a rock you hit a nazi	NO
Human	all my friends and peers are being openly racist towards asians and this bitch is not having it	OF
BERT-MRP	all my friends and peers are being openly racist towards asian ##s and this bitch is not having it	OF
BERT-RP	all my friends and peers are being openly racist towards asian ##s and this bitch is not having it	OF
BERT-HX	all my friends and peers are being openly racist towards asian ##s and this bitch is not having it	OF
Human	hey nazi dirtbag is your black bull keeping you occupied	OF
BERT-MRP	hey nazi dirt ##bag is your black bull keeping you occupied	OF
BERT-RP	dirt ##bag is your black bull keeping you occupied	OF
BERT-HX	hey nazi dirt ##bag is your black bull keeping you occupied	OF
Human	white bitch pink pussy	OF
BERT-MRP	hey white bitch pink pussy	OF
BERT-RP	white bitch pink pussy	OF
BERT-HX	white bitch pink pussy	OF

Table A2: The highlighted words of the human rationale and the rationale of the models with detection results in ’offensive’ sentences.