PEACE: Cross-Platform Hate Speech Detection - A Causality-Guided Framework

Paaras Sheth[1(✉)], Tharindu Kumarage[1], Raha Moraffah[1], Aman Chadha[2,3], and Huan Liu[1]

[1] Arizona State University, Tempe, AZ, USA
{psheth5,kskumara,rmoraffa,huanliu}@asu.edu
[2] Stanford University, Stanford, CA, USA
hi@aman.ai
[3] Amazon Alexa AI, Sunnyvale, CA, USA

Abstract. Hate speech detection refers to the task of detecting hateful content that aims at denigrating an individual or a group based on their religion, gender, sexual orientation, or other characteristics. Due to the different policies of the platforms, different groups of people express hate in different ways. Furthermore, due to the lack of labeled data in some platforms it becomes challenging to build hate speech detection models. To this end, we revisit if we can learn a generalizable hate speech detection model for the cross platform setting, where we train the model on the data from one (source) platform and generalize the model across multiple (target) platforms. Existing generalization models rely on linguistic cues or auxiliary information, making them biased towards certain tags or certain kinds of words (e.g., abusive words) on the source platform and thus not applicable to the target platforms. Inspired by social and psychological theories, we endeavor to explore if there exist inherent causal cues that can be leveraged to learn generalizable representations for detecting hate speech across these distribution shifts. To this end, we propose a causality-guided framework, **PEACE**, that identifies and leverages two intrinsic causal cues omnipresent in hateful content: the overall sentiment and the aggression in the text. We conduct extensive experiments across multiple platforms (representing the distribution shift) showing if causal cues can help cross-platform generalization.

Keywords: Causal Inference · Generalizability · Hate-Speech Detection

1 Introduction

Warning: *this paper contains contents that may be offensive or upsetting.*

Social media sites have served as global platforms for users to express and freely share their opinions. However, some people utilize these platforms to share hateful content targeted toward other individuals or groups based on their religion, gender, or other characteristics resulting in the generation and spread of hate speech. Failing to moderate online hate speech has shown to have negative impacts in real world scenarios, ranging from mass lynchings to global increase in violence toward minorities [19].

P. Sheth and T. Kumarage—Both authors contributed equally.

A. Chadha—Work does not relate to the position at Amazon.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14169, pp. 559–575, 2023.
https://doi.org/10.1007/978-3-031-43412-9_33

Thus, building hate speech detection models has become a necessity to limit the spread of hatred. Recent years have witnessed the development of these models across disciplines [2,13,27,39].

Hate speech varies based on the platform and the specific targets of the speech, influenced by factors such as social norms, cultural practices, and legal frameworks. Platforms with strict regulation policies may lead to users expressing hate in subtle ways (e.g., sarcasm), while platforms with lenient policies may have more explicit language. Collecting large labeled datasets for hate speech detection models is challenging due to the emotional burden of labeling and the requirement for skilled annotators [21]. One solution is to train a generalizable model under a cross-platform setting, leveraging the labeled data from other platforms.

Recent works developed to improve the cross-platform performance utilize either linguistic cues such as vocabulary [29] or Parts-Of-Speech (POS) tags [22]. Another direction leverages datasets with auxiliary information such as implications of various hate posts [17] or the groups or individuals attacked in the hate post [15]. Although effective, these methods suffer from shortcomings, such as linguistic methods form spurious correlations towards certain POS tags (e.g., adjectives and adverbs) or a particular category of words (e.g., abusive words). In addition, methods that utilize auxiliary information (e.g., implications of the post or the target(s)) are not extendable as the auxiliary information may not be available for large datasets or different platforms.

In contrast to previous approaches, we contend that identifying inherent causal cues is necessary for developing effective cross-platform hate speech detection models that can distinguish between hateful and non-hateful content. Since causal cues are immune to distribution shifts [5], leveraging them for learning the representations can aid in better generalization. Various studies in social sciences and psychology verify the existence of several cues that can aid in detecting hate [4,9,18,34,44] such as the hater's prior history, the conversational thread, overall sentiment, and aggression in the text. However, when dealing with a cross-platform setting, several cues may not be accessible. For instance, not all platforms allow access to user history or the entire conversation thread. Thus, we propose to leverage two causal cues namely, the overall sentiment and the aggression in the text. Both these cues can be measured easily with the aid of aggression detection tasks [3] and sentiment analysis task [43]. Moreover, both aggression and sentiment are tightly linked to hate speech. For instance, due to the anonymity on online platforms, users adopt more aggressive behavior when expressing hatred towards someone [31]. Thus, the aggression in the content could act as a causal cue to indicate hate. Similarly, hateful content is meant to denigrate someone. Thus, the sentiment also serves as a causal cue [30].

To this end, we propose a novel causality-guided framework, namely, **P**latform-ind**E**pendent c**A**usal **C**ues for generalizable hat**E** speech detection **PEACE**[1], that leverages the overall sentiment and the aggression in the text, to learn generalizable representations for hate speech detection across different platforms. We summarize our main contributions as follows:

– We identify two causal cues, namely, the overall sentiment and the aggression in the text content, to learn generalizable representations for hate speech detection.

[1] The code for PEACE can be accessed from: https://github.com/paras2612/PEACE.

– We propose a novel framework, namely, **PEACE** consisting of multiple modules to capture the essential latent features helpful for predicting sentiment and aggression. Finally, we utilize these features and the original content to learn generalizable representations for hate speech detection.
– Experimental results on five different platforms demonstrate that **PEACE** achieves state-of-the-art performance compared with vital baselines, and further experiments highlight the importance of each causal cue and interpretability of **PEACE**.

2 Related Work

Social media provides a vast and diverse medium for users to interact with each other effectively and share their opinions. Unfortunately, however, a large share of users exploits these platforms to spread and share hateful content mainly directed toward an individual or a group of people. Considering the massive volume of online posts, it is impractical to moderate them manually. To address this shortcoming, researchers have proposed various methods ranging from lexical-based approaches [14,22,38] to deep learning-based approaches [24,32,36].

However, these models have been shown to possess poor generalization capabilities. Hate speech on social media is highly volatile and is constantly evolving. A hate speech detection model that fails to generalize well may exhibit poor detection skills when dealing with a new topic of hate [10,26] or when dealing with different styles of expressing hate [1,8], thus making it critical to develop generalizable hate speech detection models. Over recent years there has been an increase in developing generalizable models.

Generalizable hate speech detection methods can be broadly classified into two parts, namely models that leverage auxiliary information such as implications of hate posts [17], information of the dataset annotators [41], or user attributes [36]. For instance, the authors of the work [17] proposed a generalizable model for implicit hate speech detection that utilizes the implications of hateful posts and learns contrastive pairs for a more generalizable representation of the hate content. Similarly, the authors of the work [41] argue that when dealing with subjective tasks such as hate speech detection, it is hard to achieve agreement amongst annotators. To this end, they propose leveraging the annotator's characteristics and the ground truth label during the training to learn better representations and improve hate speech detection. Unlike annotators' information, the authors of [36] trained a bert model with users' profiles and related social environment and generated tweets to infer better representations for hate speech detection. Although these models have improved generalizability, the auxiliary information utilized may not be easily accessible and challenging to get when dealing with cross-platform settings.

Since language models are trained on large corpora, they exhibit some generalization prowess [35]. However, the generalization can be improved by finetuning these models on datasets related to a specific downstream task. Thus, the second category leverages language models such as BERT [11] and finetuning them on large hate speech corpora [6,23]. For instance, the authors of [6] finetuned a BERT model on approximately 1.6 million hateful data points from Reddit and generated HateBERT, a state-of-the-art model for hate speech detection. Similarly, the authors of [23] finetuned BERT

for explainable hate speech detection. Aside from these works, some methods focus on leveraging lexical cues such as vocabulary used [33], emotion words, and different POS tags in the content [22], the target-specific keyphrases [12].

Although these methods have been shown to improve hate speech detection capabilities, these require large labeled corpora for finetuning language models, which may not be feasible in the real-world setting as the number of posts generated in a moment is extremely large or rely on lexical features which may not aid as a lot of the social media posts are filled with grammatical inconsistencies (such as misspelled words). In this work, inspired by works in social and psychological fields, we leverage inherent characteristics readily available in the text to learn generalizable representations, such as the aggression and the overall sentiment of the text.

3 Methodology

This section describes the methodology behind our **PEACE** framework. As shown in Fig. 1 the framework consists of two major components: (i) a cue extractor component and (ii) a hate detector component. The cue extractor component extracts the proposed innate cues, sentiment, and aggression. Moreover, this component is responsible for navigating the hate detector component toward learning a cross-platform generalized representation for hate speech detection. Consequently, the hate detector component classifies a given input to hate or non-hate classes while attending to the causal guidance of the cue extractor. In the subsequent sections, we discuss the cue extractor and hate detector components in detail.

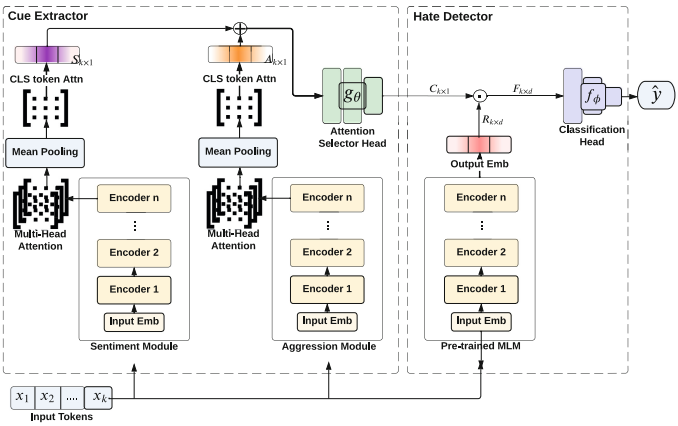


Fig. 1. Proposed framework architecture for **PEACE**. The pre-trained sentiment and aggression modules guide the representation learning process to ensure generalizability.

3.1 Causal Cue Extraction

We propose utilizing sentiment and aggression as two inherent causal cues for learning generalizable representations for better hate speech detection. Therefore, the cue extractor consists of two modules, one for extracting sentiment and one for aggression. Given an input text $X = (x_1, x_2, ..., x_k)$, the purpose of the cue extractor model is to generate an attention vector $C_{k\times1}$ where k is the input sequence length. And here, the vector $C_{k\times1}$ should represent an accumulation of sentiment and aggression score for each token in the sequence X, i.e., for a given token in the input X, $C_{k\times1}$ contains how vital that token is towards the overall input's sentiment and/or aggression. We will first discuss the architecture of each cue module (sentiment and aggression) and then elaborate on how the attention vector $C_{k\times1}$ is generated.

Sentiment Module. The sentiment module is a transformer encoder stack with n encoders that have learned a function s_γ such that given an input text $X = (x_1, x_2, ..., x_k)$, it can classify the sentiment of X, i.e., this module is a pre-trained transformer-based large language model finetuned for the sentiment detection downstream task where given an input text X, it predicts the sentiment label y (positive, neutral, negative), $y = s_\gamma(X)$.

Aggression Module. Similarly, the aggression module is also a transformer encoder stack with n encoders that have learned a function a_λ such that given an input text $X = (x_1, x_2, ..., x_k)$, it can classify whether X contains aggressive speech, i.e., this module is a pre-trained transformer-based large language model finetuned for the aggression detection downstream task where given an input text X, it predicts the aggression label y (aggressive, non-aggressive), $y = a_\lambda(X)$.

And it is essential to note here that the cue extraction module's wights are frozen when we conduct the end-to-end training of the hate detector component, i.e., we don't finetune the sentiment and aggression modules with the hate speech data.

Attention Extraction for Individual Causal Cues. As mentioned above, the cue extractor component aims to integrate the two cue modules, sentiment, and aggression, towards generating the final causal cue guidance as an attention vector $C_{k\times1}$. The first step towards this objective is extracting each individual attention vector from the cue modules. Since both the sentiment and aggression cue modules are same-sized transformer encoder stacks (n-encoders), the attention extraction process is the same for both modules. Let's take the sentiment cue module; it contains n-encoder blocks and thus consists of n multi-head attention layers. The multi-head attention layer of a given encoder block can be defined as the Eq. 1.

$$\begin{aligned} MultiHead(Q,K,V) &= head_1(Q,K,V) \oplus ...head_n(Q,K,V) \\ where; \quad head_i(Q,K,V) &= softmax(\frac{QK^T}{\sqrt{d_i}})V \end{aligned} \tag{1}$$

Here Q, K, V are Query, Key, and Value vectors of the transformer block i, and d_i is the hidden state size [37].

Our goal in using the sentiment cue module attention is to figure out the words/phrases in the input text that has particular importance towards the sentiment of the text. Therefore, we need to consider an encoder block that gives comprehensive attention to the whole input. Previous research shows that the attention heads in the BERT model's last encoder block have very broad attention - i.e., attending broadly to the entire input [7]. The architecture we consider for the sentiment module is similar to the BERT architecture (transformer encoder blocks); thus, we select the last (n^{th}) encoder block's multi-head attention layer as the candidate to extract the final attention from the sentiment module. We take the mean pooling output of the n^{th} block's multi-headed attention layer as a matrix $M_{k\times k}$ where k is the input sequence length.

$$M_{k\times k} = Mean(MultiHead_n(Q, K, V)) \tag{2}$$

Then the final attention vector $S_{k\times 1}$ for the input sequence is taken by selecting the attention at CLS token of the matrix $M_{k\times k}$. Following the same process, we extract the aggression attention vector $A_{k\times 1}$ from the aggression cue module.

Cue Integration. The final step towards creating the attention vector $C_{k\times 1}$ is to aggregate each attention vector we get from cue modules. i.e., we need to weigh and aggregate the token attentions from each cue module to get the final accumulated attention vector $C_{k\times 1}$. Once the representative attention vectors from both sentiment and aggression modules are extracted, we input the concatenated vectors through the attention selector head (g_θ). The attention selector head is a fully connected neural network that takes concatenated aggression and sentiment attention to map the final attention vector $C_{k\times 1}$.

$$C_{k\times 1} = g_\theta([S_{k\times 1} \oplus A_{k\times 1}]) \tag{3}$$

The intuition behind the attention selector head is that we need our framework to learn how to weigh the sentiment and aggression cues relevant to the context of the given input. For example, there can be cases where aggression could be the stronger cue towards hate speech than sentiment or vice versa.

3.2 Hate Detector

The hate detector component consists of a similar transformer encoder stack to learn the semantic representation of the given input. However, the output of the cue detector component, attention vector $C_{k\times 1}$, will be provided as an auxiliary signal. We select the representation learned by the hate detector blocks as $R_{k\times d}$ where k is the sequence length, and d is the hidden state size of an encoder block. Then the extracted attention is used to navigate the hate detector to adjust the representation to incorporate the causal cues. The final representation $F_{k\times d}$ is calculated as; $F_{k\times d} = R_{k\times d} \odot C_{k\times 1}$. Then the representation corresponding to the end of the sequence token ($F^{CLS}_{1\times d}$) is passed through the classification head (f_ϕ). The classification head (f_ϕ) is a fully connected neural network that takes the learned semantic embedding as the input and predicts the hate label $\hat{y}$ as $\hat{y} = f_\phi(F^{CLS}_{1\times d})$. The overall framework is trained via the cross-entropy loss for the classification, where y is the ground truth.

$$L = -\sum_{i} y_i \log(\hat{y}_i) \tag{4}$$

4 Experiments

This section discusses the experimental settings used to validate our framework, including the datasets and evaluation metrics used, and the baselines, followed by a detailed analysis of the experiments. We conducted a series of experiments to understand whether the identified causal cues, namely the sentiment and the aggression in the text, can aid in learning generalizable representations for hate speech detection and answer the following research questions.

- **RQ.1** Does the identified causal cues, namely, sentiment and aggression, enhance the generalization performance?
- **RQ.2** What is the importance of each causal cue in improving the generalization performance (ablation study)?
- **RQ.3** Which features does the **PEACE** utilize in input and whether these features are causal when compared to the other baselines?

Table 1. Dataset statistics of the experimental datasets with corresponding platforms and percentage of hateful comments or posts.

Datasets	Description	Number of Posts/Comments	Hateful Posts/Comments	Percent of Hateful Posts/Comments
GAB [15]	A collection of posts from the GAB social media platform	31,640	7,657	24.2
Reddit [28]	Conversation threads from the Reddit platform	13,633	4,219	31
Wikipedia [40]	A collection of comments on Wikipedia website	1,13,728	22,796	20
Twi-Red-You	Social media comments from three sites, namely, Twitter, Reddit, and YouTube	86,283	49,273	57.2
FRENK	Social media comments from Facebook targeting LGBT and Migrants	10,034	3,592	35.8

4.1 Datasets and Evaluation Metrics

We perform binary classification of detecting hate speech on various widely used benchmark hate datasets. Since we aim to verify cross-platform generalization, for cross-platform evaluation, we use four datasets from different platforms: Wikipedia, Facebook, Reddit, GAB, and Twitter-Reddit-YouTube. All datasets are in the English language. Wikipedia dataset [40] is a collection of user comments from the Wikipedia platform consisting of binary labels denoting whether a comment is hateful. Reddit [28] is a collection of conversation threads classified into hate and not hate. GAB [15] is a

collection of annotated posts from the GAB website. It consists of binary labels indicating whether a post is hateful or not. Finally, Twitter-Reddit-YouTube [16] is a collection of posts and comments from three platforms: Twitter, Reddit, and YouTube. It contains ten ordinal labels (sentiment, (dis)respect, insult, humiliation, inferior status, violence, dehumanization, genocide, attack/defense, hate speech), which are debiased and aggregated into a continuous hate speech severity score (hate speech score). We binarize this data such that any data with a hate speech score less than 0.5 is considered non-hateful and vice-versa. Although Twi-Red-You and Reddit both contain data from Reddit, these data do not necessarily have the same distribution. The distribution of datasets from the same platform can still defer due to variations in the timestamps, targets, locations, and demographic attributes. The FRENK dataset [20] contains Facebook comments in English and Slovene covering LGBTQ and Migrant targets. We only consider the English dataset. The dataset was manually annotated for different types of unacceptable discourses (e.g., violence, threat). We use the binary hate speech classes hate and not-hate. A summary of the datasets can be found in Table 1. For comparison with baseline methods, macro F-measure (F1) is used as an evaluation metric for validation.

4.2 Baselines

- **ImpCon (AugCon Variant)** [17] - this baseline utilizes contrastive learning with data augmentation to map similar posts closer to each other in the representation space to enable better generalization.
- **POS+EMO** [22] - this baseline proposed to use linguistic cues such as POS tags, stylometric features, and emotional cues derived by different words and the global emotion lexicon named, NRC lexicon [25] to enhance the generalizable capabilities for multilingual cross-domain hate speech detection.
- **HateBERT** [6] - finetune the BERT-base model using approximately 1.5 million Reddit messages published by suspended communities for promoting hateful content. It results in a shifted BERT model that has learned language variety and hate polarity (e.g., hate, abuse). We report the results of fine-tuned HateBERT for all the datasets.
- **HateXplain** [23] - fine-tuned using hate speech detection datasets from Twitter and Gab for a three-class classification task (hate, offensive, or normal). It combines human-annotated rationales and BERT to improve performance by reducing unintended bias toward target communities. For each dataset, we present the results of fine-tuned HateXplain.

Both HateBERT and HateXplain are not explicitly designed for generalizability but primarily for better hate speech detection. We include these baselines as they are state-of-the-art hate speech detection methods, and due to the generalization capabilities of large language models these baselines do possess better generalization [17,42].

4.3 Implementation Details

Our framework **PEACE** is built using the Huggingface Transformers library. We utilized existing RoBERTa-base models that were finetuned on social media posts for sentiment and aggression detection tasks. Additionally, a pre-trained RoBERTa-base model with 12 encoder blocks was used for the hate detection module.

During training, we employed cross-entropy loss with class balancing and optimized the framework using the Adam optimizer. The learning rate was set to 0.00002, and a dropout rate of 0.2 was used for optimal performance. Training was conducted on a NVIDIA GeForce RTX 3090 GPU with 40 GB VRAM, and the early-stopping strategy was employed.

Table 2. Cross-platform and in-dataset evaluation results for the different baseline models compared against **PEACE**. Boldfaced values denote the best performance and the underline denotes the second-best performance among different baselines.

Platforms		HateBERT	ImpCon (AugCon variant)	HateXplain	POS+EMO	PEACE
Source	Target					
Twi-Red-You	**GAB**	0.58	0.58	0.60	0.54	**0.63**
	Reddit	0.71	0.64	**0.74**	0.54	**0.74**
	Wikipedia	0.71	0.70	0.70	0.60	**0.78**
	Twi-Red-You	**0.96**	0.94	0.92	0.87	0.95
	FRENK	0.46	0.44	0.48	0.45	**0.53**
GAB	**GAB**	**0.84**	0.65	**0.84**	0.76	0.76
	Reddit	0.69	0.64	0.70	0.56	**0.71**
	Wikipedia	0.74	0.64	0.70	0.49	**0.78**
	Twi-Red-You	0.61	**0.71**	0.61	0.59	0.70
	FRENK	**0.71**	0.57	0.60	0.59	0.69
Reddit	**GAB**	0.56	0.51	0.59	0.53	**0.61**
	Reddit	0.88	0.84	**0.89**	0.59	0.88
	Wikipedia	0.66	0.63	0.64	0.56	**0.74**
	Twi-Red-You	0.73	0.70	0.77	0.65	**0.78**
	FRENK	0.42	0.42	0.44	0.49	**0.54**
Wikipedia	**GAB**	0.65	0.63	0.64	0.56	**0.68**
	Reddit	0.73	0.71	**0.74**	0.58	0.72
	Wikipedia	0.95	0.93	0.86	0.94	**0.97**
	Twi-Red-You	0.73	0.72	0.74	0.69	**0.78**
	FRENK	0.60	0.51	0.61	0.52	**0.65**
FRENK	**GAB**	0.65	0.67	0.63	0.58	**0.69**
	Reddit	0.62	0.66	0.66	0.55	**0.71**
	Wikipedia	0.67	0.76	0.73	0.53	**0.81**
	Twi-Red-You	0.65	0.65	0.64	0.62	**0.78**
	FRENK	0.78	**0.79**	0.75	0.72	0.78

4.4 RQ.1 Performance Comparison

Cross-Platform Generalization. We compare the different baseline models with **PEACE** on five real-world datasets. To evaluate the generalization capabilities of the models for each dataset, we split the data into train and test tests. We train all the models on the training data for one platform and evaluate the test sets of all the platforms. Table 2 demonstrates the performance comparison across the different test sets for the macro-F1 metric. The column **Platforms** showcases the Source platform on which the models were trained and the Target platforms used for evaluation. For each source dataset, we show the Average Performance of each model in both in-platform and cross-platform settings. As a result, we have the following observations regarding the cross-platform performance w.r.t. RQ.1:

- Overall, **PEACE** consistently yields the best performance across cross-platform evaluation for all the datasets while maintaining good in-platform macro F1. Comparing only the cross-platform performance, **PEACE** leads to a 5% improvement when trained on the Twi-Red-You dataset, 3% improvement for the GAB dataset, 6% improvement for Reddit, 3% improvement for the Wikipedia dataset, and 4% improvement for FRENK dataset.
- Among the four baselines, HateBERT serves as the strongest baseline in most cases, followed by HateXplain. This result is justified as both HateBERT and HateXplain are fine-tuned BERT models on large corpora of hateful content. We further fine-tune both HateBERT and HateXplain for each dataset. ImpCon performs well for some of the combinations, while for others, it cannot outperform HateBERT and HateXplain. We believe this is because the AugCon variant utilizes simple data augmentation. As a result, it might not be able to learn as good representations as the ImpCon variant that leverages the implications of hate. Furthermore, the utilization of the ImpCon variant is a challenging task in real-world scenarios, as the implications are not readily available for large datasets.
- The linguistic feature-based baseline (POS + EMO) doesn't generalize well to these datasets. We argue this is because the posts in these datasets are highly unstructured and grammatically incorrect. Even after pre-processing the inferred POS tags and emotion words may not be reflective of the hate content. As a result, the reliance on these features hurts the generalization performance.
- Majority of the baselines attain improved performance when trained on the Wikipedia dataset. We argue this is because of the size of the dataset. Among the four datasets, Wikipedia is the largest dataset indicating that a model can generalize better when it's trained on large datasets.

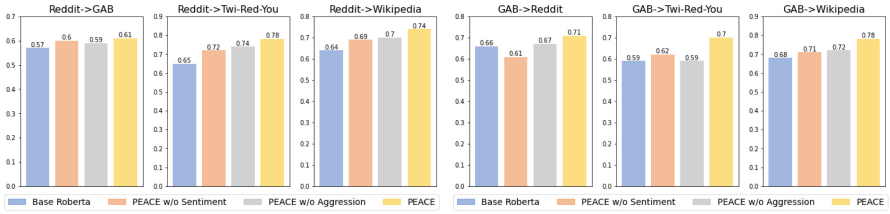
Cross-Target Generalization. Furthermore, we also conducted another experiment for the FRENK dataset to evaluate how the different models generalize in a cross-target setting, where the datasets belong to the same platform (i.e., have similar ways of expressing hate) but discuss different targets of hate. Along with the hate labels, the FRENK dataset also provides the targets of hate in the dataset, namely, *LGBTQ* and *Migrants*. Table 3 demonstrates the performance comparison for the macro-F1 metric.

Table 3. Cross-target evaluation results for the different baseline models compared against **PEACE**. Boldfaced values denote the best performance among different baselines.

targets		HateBERT	ImpCon (AugCon variant)	HateXplain	POS+EMO	PEACE
Source	Target					
Migrants	**LGBTQ**	0.74	0.68	0.65	0.61	**0.78**
LGBTQ	**Migrants**	0.66	0.67	0.64	0.58	**0.72**

We had the following observations regarding the cross-target generalization performance w.r.t. **RQ.1**:

- Comparing the cross-target generalization, we observe that C-Hate leads to an average gain of 4% improvement over the baselines. The results indicate that utilizing causal cues such as the overall sentiment and the aggression aids in learning generalizable representations and improve cross-target generalization performance.
- Across the different baselines HateBERT and ImpCon perform the best. The overall performance of HateBERT indicate that the large language models such as BERT when fine-tuned on a particular downstream task (fine-tuning BERT on hate content resulted in generation of HateBERT) can lead to competitive generalization capabilities. Furthermore, the ImpCon model performs well as it leverages data augmentation which results in more training data leading to better generalization.



(a) Reddit (b) GAB

Fig. 2. Comparison of cross-platform macro-F1 score to calculate the importance of each cue compared with the final model for Reddit and GAB datasets.

4.5 RQ.2 Importance of Each Cue

To assess the individual importance of the different causal cues used in **PEACE** with regard to the performance, we conduct the following experiments. We consider three variants of **PEACE**, one which utilizes only sentiment as the causal cue, namely, *Sentiment* one which utilizes only aggression as the causal cue, namely, *Aggression*, and one which utilizes a RoBERTa base classifier without any causal cues, namely, *Base Roberta*. We conduct cross-platform experiments by training these three variants on the Reddit and the GAB datasets. The results obtained can be seen in Fig. 2(a) for Reddit and Fig. 2(b) for GAB. As observed, **PEACE** performs the best when both causal cues

are considered. The results can deteriorate by as little as 5% to as high as 13% without the inclusion of causal cues. Among the three variants, it is observed that **PEACE** mostly benefits from the aggression cue and for some datasets, it benefits from the sentiment cue. The main reason for aggression being a strong cue is because aggression and hate are very similar tasks and earlier works have shown that aggression leads to hatred [34]. However, the base model consistently does worst, indicating that the utilization of causal cues is important to enhance the generalizability performance for hate speech detection.

Table 4. Case study illustrating the different features/tokens chosen as important tokens to detect hateful content across different models. Darker shades of the color represents the importance level of the token.

Model	Platform	
	Gab	**Reddit**
HateXplain	the feminist liberal sheeple , will finally get it one day very soon	you ' re subscribed to the christianity subreddit for reasons other than demanding the forced eradication of christian putridity . that in itself makes you unworthy .
ImpCon	the feminist liberal sheeple , will finally get it one day very soon	you ' re subscribed to the christianity subreddit for reasons other than demanding the forced eradication of christian putridity . that in itself makes you unworthy .
Sentiment	The feminist liberal sheeple, will finally get it one day very soon	You're subscribed to the Christianity subreddit for reasons other than demanding the forced eradication of Christian putridity. That in itself makes you unworthy.
+	+	+
Ours Aggression	The feminist liberal sheeple, will finally get it one day very soon	You're subscribed to the Christianity subreddit for reasons other than demanding the forced eradication of Christian putridity. That in itself makes you unworthy.
↓	↓	↓
Full Model	The feminist liberal sheeple, will finally get it one day very soon	You're subscribed to the Christianity subreddit for reasons other than demanding the forced eradication of Christian putridity. That in itself makes you unworthy.

4.6 RQ.3 Case Study

Here we provide a case study that verifies the importance of causal cues in identifying the correct context for detecting hate speech Moreover, here we visually compare **PEACE** token level attention with the baseline models HateXplain and ImpCon. In order to visualize the token importance of a given model towards its prediction, we followed a similar procedure as the cue extractor [7], where the final encoder block's attention layer was utilized to accumulate the token importance by visualizing the attention weights.

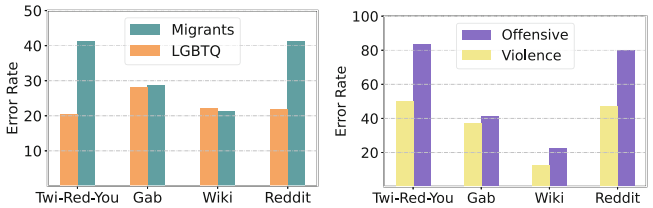
We randomly sampled hate speech text from Reddit and Gab platforms to select candidate examples for the case study. Table 4 shows a few such samples with the attention token importance visualization. In the **C-Hate's** row, we annotate the sentiment module attention in violet and aggression module attention orange. The example from the Gab platform is an instance of hate towards feminist liberals. The word *"sheeple"* and phrase *"get it one day"* can be considered as the deciding components of the text being hate speech. In contrast to the HateXplain and ImpCon, **PEACE** is attending to the word "sheeple" correctly. And we see that both the sentiment and aggression modules are giving high importance to the *"sheeple."* We have a similar observation about

the phrase *"get it one day"* where **PEACE** is successful in giving more attention to that phrase towards hate speech detection. A notable observation here is that the sentiment module is attending to the above phrase well, which could be the reason behind **C-Hate's** successfully identifying the correct context towards hate.

The next example from the Reddit platform was a complex sentence for hate speech detection, given that hate is implied, not directly expressed. As we can see, both ImpCon and HateXplain models tend to the word *"putridity"* but not to the critical contextual components that signify implicit hate, such as *"forced eradication"* and *"unworthy."* This example illustrates the issue in vocabulary-based approaches to generalized hate speech detection. On the contrary, we can see that the sentiment and aggression modules accurately attend to the *"forced eradication"* and *"unworthy"* phrases navigating **PEACE** to correctly identify the hate speech context.

5 Limitations and Error Analysis

In this section, we conduct an error analysis to better understand our work's limitations and aid future work in cross-platform generalized hate speech detection. For this analysis, we select the FRENK dataset (Facebook) as the testing dataset, given it contains fine-grained information about the data, such as hate targets (LGBTQ vs. migrants) and hate types (offense vs. violence). We used the **PEACE** models trained on other platforms (Twitter, Gab, Reddit, and Wiki) to run the test on the FRENK dataset mentioned above. Finally, we analyze each model's misclassification rate/error rate under dimensions of hate target and hate type.



(a) target (b) Type

Fig. 3. Analysing error rate of **PEACE** under different Dimensions such as (a) hate targets (LGBTQ vs. migrants) and (b) hate type (offense vs. violence).

As seen in Fig. 3(a), the model tends to have a higher error rate in detecting migrants-related samples, particularly when trained on Reddit and Twi-Red-You datasets. One notable characteristic we observed in the Reddit and Twi-Red-You datasets is that the hate examples tend to include a majority of targeted hate towards particular individuals. Similarly, the LGBTQ target in FRENK dataset contains a majority of hate examples towards individuals. However, in contrast, the migrant target contains more generic hate examples towards a group of people. This mismatch in training and

Table 5. Examples representing the different kinds of hate. The violence hate type is more explicit and direct, whereas the offense hate type is more subtle and implicit.

Hate Type	Examples
Violence	shoot them all, done!!! let the communists solve the problem!!! coz i believe that these people wont stop, sooner or later, Germany will have to use guns
	Quick... Bomb it
Offensive	The annoying thing is that 75% of the migrants are Young men, why aren't they fighting for THEIR country? Or is it more a case of they can get more from European countries (money, house,education etc.)
	Are there terrorists hidden in migration groups? Likely

testing platforms might be causing the high error rate in the migrants compared to the LGBTQ.

The error analysis (Fig. 3(b)) reveals that the **PEACE** model exhibits a higher error rate in the offensive hate type compared to the violence type. To further investigate this, we examine the textual traits associated with each hate type. Representative samples from both categories are provided in Table 5. In the violence hate type, the hate aspect is explicit and easily recognizable to both readers and the model. Sentiment and aggression cues are also readily detectable in these instances. However, in the offensive hate type, hate is inherently more implicit than explicit. Consequently, learning valuable signals through causal cues becomes challenging when the expressed hatred is implicit.

6 Conclusions and Future Work

Social media platforms facilitate global opinion sharing but are often misused for spreading targeted hate speech. Automatic hate speech detection is crucial but challenging due to evolving hate and limited labeled data. To address this, we propose **PEACE**, a hate speech detection model that leverages aggression, sentiment, and causal cues to learn generalizable representations. Our extensive experiments demonstrate that **PEACE** outperforms state-of-the-art baselines on multiple platforms and targets. We also emphasize the importance of each causal cue and perform case studies to identify the features used by **PEACE** for hate speech detection. To enhance **PEACE**'s generalization, we will explore automating the identification of causal cues and develop an end-to-end system.

Acknowledgements. This material is based upon work supported by, or in part by the Office of Naval Research (ONR) under contract/grant number N00014-21-1-4002, the Army Research Office under the grant number W911NF2110030, and Defense Advanced Research Projects Agency (DARPA) under the grant number HR001120C0123. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Ethical Statement.

Freedom of Speech and Censorship. Our research aims to develop algorithms that can effectively identify and mitigate harmful language across multiple platforms. We recognize the importance of protecting individuals from the adverse effects of hate speech and the need to balance this with upholding free speech. Content moderation is one application where our method could help censor hate speech on social media platforms such as Twitter, Facebook, Reddit, etc. However, one ethical concern is our system's false positives, i.e., if the system incorrectly flags a user's text as hate speech, it may censor legitimate free speech. Therefore, we discourage incorporating our methodology in a purely automated manner for any real-world content moderation system until and unless a human annotator works alongside the system to determine the final decision.

Use of Hate Speech Datasets. In our work, we incorporated publicly available well-established datasets. We have correctly cited the corresponding dataset papers and followed the necessary steps in utilizing those datasets in our work. We understand that the hate speech examples used in the paper are potentially harmful content that could be used for malicious activities. However, our work aims to help better investigate and help mitigate the harms of online hate. Therefore, we have assessed that the benefits of using these real-world examples to explain our work better outweigh the potential risks.

Fairness and Bias in Detection. Our work values the principles of fairness and impartiality. To reduce biases and ethical problems, we openly disclose our methodology, results, and limitations and will continue to assess and improve our system in the future.

References

1. Ali, R., Farooq, U., Arshad, U., Shahzad, W., Beg, M.O.: Hate speech detection on twitter using transfer learning. Comput. Speech Lang. **74**, 101365 (2022)
2. Alkomah, F., Ma, X.: A literature review of textual hate speech detection methods and datasets. Information **13**(6), 273 (2022)
3. Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 90–97 (2018)
4. Bauwelinck, N., Lefever, E.: Measuring the impact of sentiment for hate speech detection on Twitter. Proc. HUSO, 17–22 (2019)
5. Bühlmann, P.: Invariance, causality and robustness. Stat. Sci. (2020)
6. Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: HateBERT: retraining BERT for abusive language detection in english. arXiv preprint arXiv:2010.12472 (2020)
7. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 276–286 (2019)
8. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: Cross-platform evaluation for Italian hate speech detection. In: CLiC-it 2019–6th Annual Conference of the Italian Association for Computational Linguistics (2019)
9. Craig, K.M.: Examining hate-motivated aggression: a review of the social psychological literature on hate crimes as a distinct form of aggression. Aggress. Violent. Beh. **7**(1), 85–101 (2002)

10. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: hate speech detection on Facebook. In: Proceedings of the first Italian conference on cybersecurity (ITASEC 2017), pp. 86–95 (2017)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding, E.: Hate lingo: a target-based linguistic analysis of hate speech in social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
13. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
14. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimedia Ubiquit. Eng. **10**(4), 215–230 (2015)
15. Kennedy, B., et al.: The gab hate corpus: a collection of 27k posts annotated for hate speech. PsyArXiv. **18** (2018)
16. Kennedy, C.J., Bacon, G., Sahn, A., von Vacano, C.: Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. arXiv preprint arXiv:2009.10277 (2020)
17. Kim, Y., Park, S., Han, Y.S.: Generalizable implicit hate speech detection using contrastive learning. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 6667–6679 (2022)
18. Krahé, B.: The Social Psychology of Aggression. Routledge (2020)
19. Laub, Z.: Hate speech on social media: global comparisons. Counc. Foreign Relat. **7** (2019)
20. Ljubešić, N., Fišer, D., Erjavec, T.: The FRENK datasets of socially unacceptable discourse in Slovene and English. In: Ekštein, K. (ed.) TSD 2019. LNCS (LNAI), vol. 11697, pp. 103–114. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_9
21. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS ONE **14**(8), e0221152 (2019)
22. Markov, I., Ljubešić, N., Fišer, D., Daelemans, W.: Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 149–159 (2021)
23. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14867–14875 (2021)
24. Mazari, A.C., Boudoukhani, N., Djeffal, A.: BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput., 1–15 (2023)
25. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Comput. Intell. **29**(3), 436–465 (2013)
26. Pamungkas, E.W., Basile, V., Patti, V.: A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. Inf. Process. Manag. **58**(4), 102544 (2021)
27. Paz, M.A., Montero-Díaz, J., Moreno-Delgado, A.: Hate speech: a systematized review. SAGE Open **10**(4), 2158244020973022 (2020)
28. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251 (2019)
29. Ramponi, A., Tonelli, S.: Features or spurious artifacts? Data-centric baselines for fair and robust hate speech detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3027–3040. Association for Computational Linguistics, Seattle, United States, July 2022
30. Rodriguez, A., Argueta, C., Chen, Y.L.: Automatic detection of hate speech on Facebook using sentiment and emotion analysis. In: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 169–174. IEEE (2019)

31. Rösner, L., Krämer, N.C.: Verbal venting in the social web: effects of anonymity and group norms on aggressive language use in online comments. Soc. Media+ Soc. **2**, 2056305116664220 (2016)
32. Roy, S.G., Narayan, U., Raha, T., Abid, Z., Varma, V.: Leveraging multilingual transformers for hate speech detection. arXiv preprint arXiv:2101.03207 (2021)
33. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10 (2017)
34. Sengupta, A., Bhattacharjee, S.K., Akhtar, M.S., Chakraborty, T.: Does aggression lead to hate? Detecting and reasoning offensive traits in Hinglish code-mixed texts. Neurocomputing **488**, 598–617 (2022)
35. Tamkin, A., Singh, T., Giovanardi, D., Goodman, N.: Investigating transferability in pretrained language models. arXiv preprint arXiv:2004.14975 (2020)
36. del Valle-Cano, G., Quijano-Sánchez, L., Liberatore, F., Gómez, J.: SocialHaterBERT: a dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles. Exp. Syst. Appl. **216**, 119446 (2023)
37. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
38. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words-a feature-based approach. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1046–1056 (2018)
39. Williams, M.L., Burnap, P., Javed, A., Liu, H., Ozalp, S.: Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. Br. J. Criminol. **60**(1), 93–117 (2020)
40. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399 (2017)
41. Yin, W., Agarwal, V., Jiang, A., Zubiaga, A., Sastry, N.: AnnoBERT: effectively representing multiple annotators' label choices to improve hate speech detection. arXiv preprint arXiv:2212.10405 (2022)
42. Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Comput. Sci. **7**, e598 (2021)
43. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowl. Inf. Syst. **60**, 617–663 (2019)
44. Zhou, X., et al.: Hate speech detection based on sentiment knowledge sharing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7158–7166 (2021)