Improving Cross-Domain Hate Speech Generalizability with Emotion Knowledge

Shi Yin Hong and **Susan Gauch**
Department of Electrical Engineering and Computer Science
University of Arkansas, Fayetteville, AR, USA
{syhong, sgauch}@uark.edu

arXiv:2311.14865v2 [cs.CL] 17 Dec 2023

Abstract

Reliable automatic hate speech (HS) detection systems must adapt to the in-flow of diverse new data to curtail hate speech. However, hate speech detection systems commonly lack generalizability in identifying hate speech dissimilar to data used in training, impeding their robustness in real-world deployments. In this work, we propose a hate speech generalization framework that leverages emotion knowledge in a multitask architecture to improve the generalizability of hate speech detection in a cross-domain setting. We investigate emotion corpora with varying emotion categorical scopes to determine the best corpus scope for supplying emotion knowledge to foster generalized hate speech detection. We further assess the relationship between using pretrained Transformers models adapted for hate speech and its effect on our emotion-enriched hate speech generalization model. We perform extensive experiments on six publicly available datasets sourced from different online domains and show that our emotion-enriched HS detection generalization method demonstrates consistent generalization improvement in cross-domain evaluation, increasing generalization performance up to 18.1% and average cross-domain performance up to 8.5%, according to the F1 measure[1].

1 Introduction

Hate speech (HS) possesses justified regulatory grounds since it inflicts harm toward a targeted individual or group based on perceived characteristics (Gelber, 2021). The social obstruction imposed by the pervasive online HS thus ignites counteraction from the natural language processing (NLP) community to create machine learning-based systems to automate the HS identification process (Poletto et al., 2021; Alkomah and Ma, 2022). Despite efforts dedicated to the goal in the past decade, HS detection remains a challenging task to conquer (Fortuna et al., 2022; Wiegand et al., 2021). Notably, the lack of generalizability is a prevalent issue with current HS models (Yin and Zubiaga, 2021).

HS models that suffer from generalizability show a discrepancy in their performance across HS datasets (Wiegand et al., 2019). Such models are competitive in detecting HS on the data from the same source as the data they are trained with but show a significant performance gap when detecting HS from varied HS sources. The mainstream approach addressing the issue utilizes knowledge from the HS domain to improve HS generalization. Observations on the semantic distribution of the data (e.g., implicit or explicit HS) serve as the basis for counteractive augmentation, synthetic data generative, and sampling techniques to bridge linguistic gaps observed in HS (Ilan and Vilenchik, 2022; Arango et al., 2022; Wullach et al., 2021; Ludwig et al., 2022). Furthermore, datasets may be re-annotated, combined, or created to meet the generalization task by topics (Yoder et al., 2022; Nejadgholi et al., 2022; Toraman et al., 2022).

In real-world applications, however, it is unrealistic to conduct retrospective HS analysis with the constant inflow of new and changing data. Further, HS is the byproduct of the evolving social context, culture, and linguistic interpretation (Hilte et al., 2023), which elevates the challenge of using static criteria in assessing hate speech. HS models that cannot demonstrate robust generalizability cannot reliably carry out their high-stake social responsibility in safeguarding vulnerable groups from the multifarious HS reflected in diverse online platforms. The lack of generalizability of HS models can even unintentionally exacerbate the proliferation of online HS by allowing out-of-domain hateful speech to evade its consequences while curtailing free speech when sanctioning unhateful speech (Bianchi et al., 2022).

[1]Code and resources are available at https://github.com/sy-hong/ek-hs-generalizability

Dataset	Original Size	Domain
Founta (Founta et al., 2018)	99,799	Twitter
Kaggle (Peller, 2014)	312,737	Wikipedia
Kumar (Kumar et al., 2018)	15,000	Facebook
Offensive Reddit (Qian et al., 2019)	5,020	Reddit
Razavi (Razavi et al., 2010)	1,525	Natural Semantic Modules, Usenet
Waseem and Hovy (Waseem and Hovy, 2016)	16,907	Twitter
GoEmotions (Demszky et al., 2020)	58,009	Reddit

Table 1: Datasets used in the cross-domain cross-dataset generalization evaluation. The GoEmotions (Demszky et al., 2020) dataset supplies the auxiliary emotion knowledge in our multitask HS generalization framework.

In this work, we utilize emotion knowledge to support the generalization of HS detection. We find that utilizing emotion knowledge in addressing HS, which exhibits a greater relative conceptual variability, helps to mitigate the variance of HS language that challenges HS generalization. Specifically, we adopt the GoEmotions dataset (Demszky et al., 2020) to provide emotion information and investigate the effect of leveraging two variants of emotion corpora – the dataset's original release with 28 emotions and its Ekman emotion corpus (Ekman, 1971) equivalent – in improving the HS detection's generalizability in a multitask framework. We utilize BERT (Devlin et al., 2019) and fBERT (Sarkar et al., 2021) as the base Transformers models to evaluate their effectiveness in enhancing emotion-driven HS generalizability given their varying pre-trained corpora relatedness to HS. We assess the proposed model's performance in improving the generalizability of six popular benchmark datasets from different domains with the cross-dataset evaluation method. Our emotion-enriched HS detection generalization method demonstrates consistent cross-domain generalization binary F1 performance, increasing generalization performance up to 18.1% and average cross-domain up to 8.5%. Our main contributions are summarized below:

- We propose an emotion-integrated multitask HS generalization framework that utilizes emotion knowledge to strengthen cross-domain HS generalization.
- We study how the categorical scope of the emotion corpora – the 28-class GoEmotions (Demszky et al., 2020) corpus and the six-class Ekman emotion (Ekman, 1971) corpus – affects the generalizability of HS detection with our method.
- We evaluate the effect of the adopted Transformers base models', BERT and fBERT, varying adaptiveness to the HS domain on our cross-domain HS generalization framework.
- We perform extensive evaluations in cross-domain settings on six publicly available benchmark datasets with varied HS forms to show our model improves cross-domain HS generalization.

2 Related Works

2.1 Generalization of HS Detection

In studies of HS detection's generalization, some works aim to identify sources behind the lack of performance generalization. Fortuna et al. (2020) analyze the homogeneity of applied categories in popular public HS datasets and empirically support the lack of compatibility among the cross-dataset performance. They suggest an underlying reason is the lack of consensus on HS's subjective definitional concept, leading to varied criteria for HS categorization. Fortuna et al. (2021) reason that the low generalization of HS detectors roots to the imbalance of explicit and implicit distributions of HS across datasets and encourage HS dataset creators to identify precise categorization (e.g., sexism, racism) to endow levels of granularity in HS. Arango et al. (2022) argue for more transparency behind the user distribution of existing HS datasets to prevent spurious high performance of HS classifiers overfitted to limited users in data production. They propose to improve cross-dataset generalization by adopting countering sampling techniques addressing user overfitting.

The dominant method in addressing the generalization of HS detection considers analyzing underlying semantic and topical traits of HS datasets. Bourgeade et al. (2023) sample from six HS corpora and present a re-annotated dataset version based on topic-generic and topic-specific levels.

GE_{ek}	GE_{go}
Anger	Anger, annoyance, disapproval
Disgust	Disgust
Fear	Fear, nervousness
Joy	Admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
Sadness	Sadness, disappointment, embarrassment, grief, remorse
Surprise	Surprise, realization, confusion, curiosity

Table 2: Categorical emotion conversion between the version of the GoEmotions (Demszky et al., 2020) dataset adopted with the Ekman GE_{ek} (Ekman, 1971) corpus and its original GE_{go} corpus.

They find that adopting a mixture of topic-generic and topic-specific tweets in the model fine-tuning step enhances the generalization of HS classifiers. Nejadgholi et al. (2022) show the weakness of HS classifiers at generalizing implicit racism from topic-centric HS datasets and propose a model based on concept activation vector to improve the interpretability of the model in performing generalization. Ludwig et al. (2022) adopt unsupervised domain adaptation to improve HS models' ability to perform generalization across HS targeting toward a subset of categorized target groups encompassed by the HateXExplain dataset (Mathew et al., 2021). Wullach et al. (2021) adopt a GPT-based language model to generate synthetic HS via sequence generation using existing HS datasets as an augmentation technique to improve the quality of HS generalization.

2.2 Multitask HS Detection With Emotions

Multitask learning is a training methodology that has recently gained popularity in NLP due to its power to integrate knowledge from related tasks in modeling a target task (Zhang et al., 2023; Turcan et al., 2021). In the context of this study, emotion classification is the auxiliary task modeled jointly with the main task of binary hate speech classification to improve hate speech detection generalization leveraging the integrated emotion knowledge. Present works that consider emotion features in related studies focus more on the abusive language detection domain without considering the model's generalizability. Rajamanickam et al. (2020) investigate abusive language detection by incorporating emotion detection into MLP-based and BiLSTM-based networks with a hard-sharing multitask framework. Samghabadi et al. (2019) employ pre-trained DeepEmoji to assign textual data with relevant emotions in capturing offensive language. Plaza-del-Arco et al. (2021) and Plaza-del Arco et al. (2022) incorporated sentiment information in addressing HS domain-related tasks, offensive and abusive language detections, with emotion information. In the HS domain, Chiril et al. (2022) adopt emotion lexicons such as SenticNet and EmoSenticNet to detect hate speech with multi-targets from topic-generic datasets and conclude that the utilization of affective knowledge enhances hate speech detection categorized by targets and topics. Mnassri et al. (2023) perform hate and offensive language detection using emotion information in a multitask setting involving cross-lingual settings.

Diverging from previous works, we utilize emotion knowledge with varying categorical scopes to uplift cross-domain generalizability. We further examine Transformers models' relative domain adaptability to the HS in cross-domain generalizability.

3 Methodology

3.1 Experimental Datasets

We adopted the six datasets and the processing procedures used by Ilan and Vilenchik (2022), a recent hate speech generalization work, to assess our model's ability to improve hate speech detection generalizability in a cross-domain setting. These datasets are: Founta (Founta et al., 2018), Kaggle (Peller, 2014), Kumar (Kumar et al., 2018), Waseem and Hovy (W&H) (Waseem and Hovy, 2016), Offensive Reddit (Qian et al., 2019), and Razavi (Razavi et al., 2010). For larger datasets such as Founta, Kaggle, and Kumar, positive and negative HS samples were randomly sampled for 5,000 entries, totaling 10,000 samples per dataset. Negative samples in smaller datasets such as Razavi and W&H were downsampled to the sizes of positive samples, 482 and 795, respectively, to control the source of variance in generalization study (Swamy et al., 2019). For the Offensive Reddit dataset, the 3,230 positive samples were balanced with 3,230 negative samples from Washam (2019). User mentions, hashtags, URLs, and emojis were removed for text preprocessing. Table 1 shows an overview of the datasets and their respective domains used in our cross-domain HS

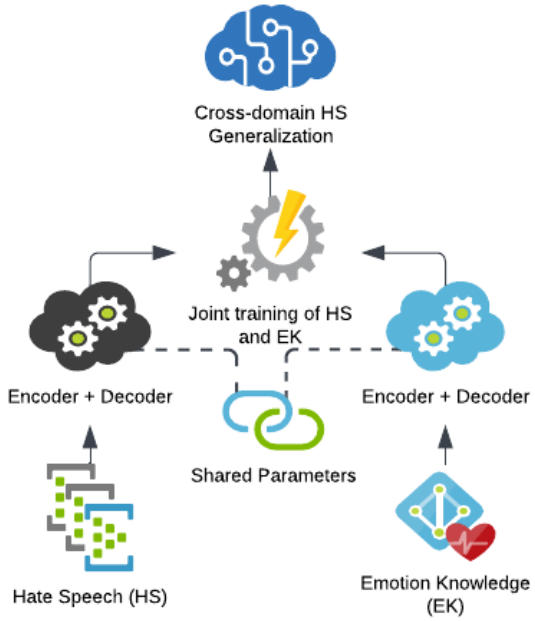

Figure 1: Abstract architecture for the emotion knowledge-enriched HS generalization framework.

generalizability study.

We use two variants of the GoEmotions dataset (Demszky et al., 2020) to supply the emotion information. The first variant is the original version of the dataset with 27 emotions plus the neutral class, which we would refer to as GE_{go} in the following sections. GE_{go} serves in our study by providing fine-grained categorical emotion information in assessing the generalizability of our emotion-enriched hate speech detection method. To contrast the detailed emotion categories supplied by GE_{go}, we convert the 28 emotion classes in GE_{go} into its Ekman equivalent with six emotions as detailed in Demszky et al. (2020). We refer to the GoEmotions dataset in the form of the traditional six-class emotion corpus as GE_{ek}. Table 2 provides the emotion categorical conversion between GE_{go} and GE_{ek}.

3.2 Emotion-enriched Hate Speech Generalization

We implement a multitask architecture as illustrated in Figure 1 given a set of disjoint tasks $\Omega = \{\Omega_{hs}, \Omega_e\}$, where Ω_{hs} denotes the main task of hate speech detection and Ω_e denotes the auxiliary task of emotion analysis. We let Ω_{hs} and Ω_e share the same Transformers-based encoding layers to promote direct knowledge transfer. This hard parameter-sharing choice prevents overfitting and overparameterization (Ruder, 2017) (Liu et al., 2019). The respective dataset D_Ω for each task is D_{hs} for Ω_{hs} and D_e for Ω_e, where $D_\Omega = \{(\chi_\Omega^{(i)}, y_\Omega^{(i)})\}|_{i=1}^N$. The input and target output are denoted by $\chi_\Omega^{(i)}$ and $y_\Omega^{(i)}$, and N is the total data entry for each task.

Given D_Ω, we first tokenize $\omega \in D_\Omega$ into their subword representations, s_ω. We obtain the embedding vector $v(s_\omega)$ via an embedding layer that transforms $[e_1(s_\omega), e_2(s_\omega), ..., e_d(s_\omega)]^T$ into their vectorial representation, where d is the dimension of the embedding space and $e_i(s_\omega)$ denotes the i^{th} element of the embedding vector. We acquire the hidden state $h_\Omega^{(i)}$ as follows:

$$h_\Omega^{(i)} = Encoder(v(s_\omega)^{(i)}, \xi_\Omega^E) \quad (1)$$

where ξ_Ω^E denotes the parameters for the encoder. We use BERT (Devlin et al., 2019) and fBERT (Sarkar et al., 2021) for encoding, where BERT is trained on a general corpus, and fBERT is trained on an offensive language corpus based on the OLID dataset (Zampieri et al., 2019). We promote the joint knowledge exchange between Ω_{hs} and Ω_e as their parameters are shared in the encoder unit. In the process, Ω_e functions as a regularizer, introducing an inductive bias as the two tasks share more general representations that make the model favors prediction $\hat{y}_{hs}^{(i)}$ that explains both tasks well.

Both tasks Ω_{hs} and Ω_e continue to share parameters at the decoding stage but are independent with separate MLP layers for each task. For each task, the predicted output $\hat{y}_\Omega^{(i)}$ is obtained as follows:

$$\hat{y}_\Omega^{(i)} = Decoder(h_\Omega^{(i)}, \xi_\Omega^D) \quad (2)$$

where ξ_Ω^D denotes the parameters of the decoder for Ω.

For single-class prediction when modeling Ω_{hs} and Ω_e using the GE_{ek} corpus, we minimize the negative log-likelihood (NLL) loss:

$$L_{NLL}(\Omega) = -\sum_{i=1}^{n}(y_\Omega^{(i)} log(\hat{y}_\Omega^{(i)}) + (1 - y_\Omega^{(i)}) log(1 - \hat{y}_\Omega^{(i)})) \quad (3)$$

where $y_\Omega^{(i)}$ denotes the ground truth label, and $\hat{y}_\Omega^{(i)}$ denotes the predicted label for Ω. For making predictions using the multi-labeled GE_{go} corpus, we apply the binary cross-entropy (BCE) loss:

$$L_{BCE}(\Omega) = -\frac{1}{N}\sum_{i=1}^{N} y_\Omega^{(i)} log(p(y_\Omega^{(i)})) + (1 - y_\Omega^{(i)}) log(1 - p(y_\Omega^{(i)})) \quad (4)$$

Train/Test	Founta	Kaggle	Kumar	Off.Red.	Razavi	W&H	CD Avg
Founta	0.922	0.800	0.470	0.734	0.672	0.543	0.669
Founta + GE_{go}	**0.926**	0.809	0.556	0.739	0.686	0.553	**0.697**
Founta + GE_{ek}	0.925	0.805	0.550	0.736	0.679	0.548	0.683
Kaggle	0.833	0.902	0.581	0.718	0.741	0.632	0.701
Kaggle + GE_{go}	0.845	0.911	0.592	0.725	0.767	0.677	0.721
Kaggle + GE_{ek}	0.839	**0.916**	0.616	0.742	0.777	0.701	**0.735**
Kumar	0.729	0.645	0.692	0.640	0.669	0.644	0.665
Kumar + GE_{go}	0.764	0.739	0.703	0.670	0.712	0.663	0.709
Kumar + GE_{ek}	0.779	0.746	**0.711**	0.680	0.741	0.664	**0.722**
Off. Red.	0.671	0.638	0.530	0.931	0.627	0.618	0.617
Off. Red. + GE_{go}	0.727	0.694	0.558	**0.932**	0.657	0.650	**0.657**
Off. Red. + GE_{ek}	0.688	0.678	0.556	0.931	0.651	0.659	0.646
Razavi	0.798	0.829	0.566	0.767	0.866	0.631	0.718
Razavi + GE_{go}	0.834	0.836	0.635	0.773	**0.890**	0.652	0.746
Razavi + GE_{ek}	0.844	0.855	0.663	0.770	0.871	0.653	**0.757**
W&H	0.535	0.500	0.518	0.659	0.545	0.896	0.555
W&H + GE_{go}	0.597	0.529	0.570	0.675	0.557	**0.899**	0.583
W&H + GE_{ek}	0.603	0.541	0.555	0.684	0.584	0.898	**0.591**

Table 3: BERT-based in-domain and cross-domain HS generalization performance

where N is the training entry count, $y_{\Omega}^{(i)}$ denotes the ground truth label and $p(y_{\Omega}^{(i)})$ denotes the prediction probability for true positive prediction for Ω.

4 Cross-Domain Generalization

4.1 Experimental Setup and Implementation Details

We compare the performance of our approach with the uncased base version of BERT and fBERT fine-tuned only with HS datasets as baselines. For our emotion-integrated HS generalization model, we adopt BERT and fBERT as the base models as shown in Table 3 and Table 4, respectively. Emotion-integrated models are noted with their respective emotion corpus, $+GE_{go}$ or $+GE_{ek}$. We perform training on one dataset and evaluation on all datasets' separate testing sets. This includes in-dataset evaluation as we perform training and testing on the same dataset, which also assesses the in-domain generalizability. We assess cross-domain generalizability performance between different training and test sets not from the same domain. We further analyze the overall generalizability performance by providing the average cross-domain binary F1 for individual experiments as shown in the last column (CD Avg) of Table 3 and Table 4.

All models were implemented using PyTorch (Paszke et al., 2019), and all experiments were conducted on NVIDIA Quadro RTX 4000. We trained all models with 5 epochs with early stopping as we often observed that the best validation performance is obtained in the first three epochs. We employed a batch size of 8 and an Adam optimizer with 1E-4 as the learning rate. The average binary F1 performance of three separate trails using seeds {0, 1, 3} are reported. The best in-domain and cross-domain average scores are in bold. Results that show improvement from baselines are highlighted in gradients of purple for in-domain settings and blue for cross-domain settings based on their relative strength of improvement.

4.2 Results

Table 3 shows the performance of our evaluation using BERT as the base model. From the baseline model, we observe a general decline in performance when models are evaluated in a cross-domain setting compared to an in-domain setting, which supports our motivation to improve cross-domain generalizability. The disparity in performance is the greatest for the Offensive Reddit, W&H, and Kaggle datasets, which show a cross-domain performance decline of 33.7% ($\frac{0.617 \quad 0.931}{0.931} \times 100$), 22.4%, and 22.2%, respectively, when average cross-domain performance is compared with the respective in-domain perfor-

Train/Test	Founta	Kaggle	Kumar	Off.Red.	Razavi	W&H	CD Avg
Founta	0.929	0.7961	0.377	0.743	0.600	0.526	0.629
Founta + GE_{go}	**0.930**	0.805	0.397	0.753	0.609	0.560	0.625
Founta + GE_{ek}	0.929	0.841	0.411	0.751	0.662	0.576	**0.666**
Kaggle	0.848	0.922	0.596	0.745	0.765	0.756	0.742
Kaggle + GE_{go}	0.854	0.925	0.609	0.754	0.775	0.769	0.752
Kaggle + GE_{ek}	0.861	**0.926**	0.618	0.757	0.785	0.783	**0.761**
Kumar	0.848	0.596	0.715	0.596	0.765	0.756	0.712
Kumar + GE_{go}	0.868	0.619	0.721	0.611	0.797	0.787	**0.736**
Kumar + GE_{ek}	0.857	0.602	**0.733**	0.600	0.797	0.787	0.729
Off. Red.	0.644	0.631	0.537	0.936	0.621	0.659	0.618
Off. Red. + GE_{go}	0.682	0.688	0.550	0.933	0.665	0.683	**0.653**
Off. Red. + GE_{ek}	0.652	0.645	0.553	**0.938**	0.665	0.700	0.643
Razavi	0.845	0.871	0.642	0.768	0.881	0.790	0.783
Razavi + GE_{go}	0.877	0.879	0.643	0.771	0.878	0.791	**0.792**
Razavi + GE_{ek}	0.859	0.872	0.644	0.769	**0.885**	0.795	0.788
W&H	0.607	0.605	0.505	0.729	0.526	0.894	0.526
W&H + GE_{go}	0.663	0.634	0.574	0.742	0.607	0.896	0.638
W&H + GE_{ek}	0.630	0.630	0.577	0.753	0.603	**0.903**	**0.640**

Table 4: fBERT-based in-domain and cross-domain HS generalization performance

mance for each evaluation dataset.

When we include emotion knowledge using the original GoEmotions dataset with its GE_{go} corpus, the performance of in-domain and cross-domain generalization improves. The greatest in-domain improvement ($\uparrow 4.5\% = \frac{\frac{0.597+0.899}{2} \quad \frac{0.535+0.896}{2}}{\frac{0.535+0.896}{2}} \times 100$) is with the W&H dataset. Adding the emotion knowledge from the GE_{go} corpus leads to an average increase in cross-domain performance from 2.9% ($\frac{0.721 \quad 0.701}{0.701} \times 100$) to 6.6% ($\frac{0.709 \quad 0.665}{0.665} \times 100$) observed in the Kaggle and Kumar datasets, respectively. Experiments that exhibit higher average out-domain improvement used Kumar ($\uparrow 6.6\%$), Offensive Reddit ($\uparrow 6.6\%$), and W&H ($\uparrow 4.9\%$) as training sets. The greatest generalizability uplifts for individual cross-domain experiments are shown in Founta $\rightarrow$ Kumar (i.e., Founta dataset is the training set for the model that generalizes on the testing set from the Kumar dataset), Kumar $\rightarrow$ Kaggle, and Razavi $\rightarrow$ Kumar experiments, resulting in generalizability enhancement of 18.1%, 14.5%, and 12.2%, respectively.

When we adopt the GoEmotions dataset based on the GE_{ek} corpus to supply emotion knowledge in our HS generalization model, both in-domain and cross-domain performances also show improvements. The in-domain performance increases up to 4.9% with the maximum uplifts corresponding to the W&H datasets. In cross-domain experiments, the average increase in binary F1 ranges from 2.1% to 8.5% with the least and best cross-domain performance average corresponding to experimental settings where Founta and Kumar datasets are used as training sets. Integrating emotion knowledge via the GE_{ek} corpus also shows competitive cross-domain generalizability when W&H ($\uparrow 6.4\%$) and Kaggle ($\uparrow 4.8\%$) datasets. The greatest generalizability uplifts for individual cross-domain experiments are shown in Razavi $\rightarrow$ Kumar ($\uparrow 17.1\%$), Kumar $\rightarrow$ Kaggle ($\uparrow 15.7\%$), and Kaggle $\rightarrow$ W&H ($\uparrow 10.9\%$) experiments. Experiments Razavi $\rightarrow$ Kumar and Kumar $\rightarrow$ Kaggle are also the top-performance individual cross-domain experiments using the GE_{go} corpus.

Table 4 shows the performance of our evaluation when we used fBERT as the base model. We also observe a general decline in performance when models are evaluated in cross-domain settings compared to in-domain settings with the baseline model's performance. The same three datasets when using the BERT as the baseline show the greatest difference in in-domain and cross-domain, resulting in a performance decline of 33.9%, 19.6%, and 16.0% for the Offensive Reddit, Kaggle, and W&H datasets, respectively.

Using the GE_{go} corpus to induce emotion knowledge in our model, we consistently observe cross-domain generalizability enhancement but not al-

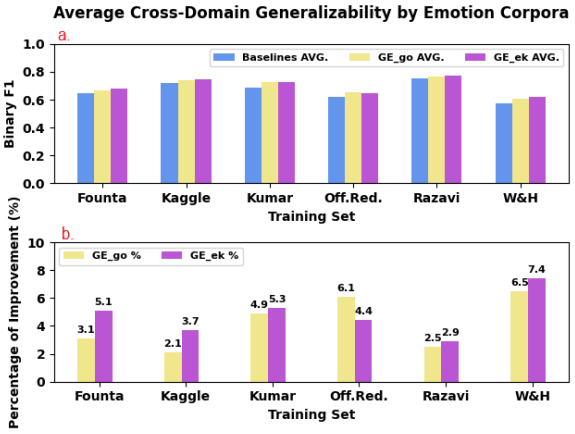

Figure 2: **a.** Average binary F1 cross-domain generalizability performance of the baseline average, GE_{go}-based model average, and GE_{ek}-based model average. **b.** Percentage uplifts compared to the baseline average of the GE_{go}-based model average and and GE_{ek}-based model average.

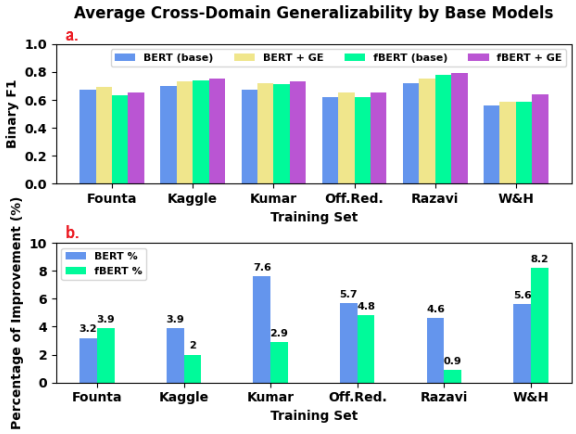

Figure 3: **a.** Average binary F1 cross-domain generalizability performance of the baseline average and emotion-enriched model average of the BERT-based and fBERT-based models. **b.** Percentage uplifts compared to the baseline average of the BERT-based model average and fBERT-based model average.

ways with in-domain experiments. For the Offensive Reddit and Razavi datasets, the in-domain performance decreases by 0.003 in binary F1. The greatest in-domain improvement (↑3.8%) is observed from the W&H dataset. The generalizability improvement in average cross-domain performance ranges from 1.2% to 8.1%. The minimum and maximum average cross-domain performances correspond to Razavi and W&H datasets, respectively. Emotion knowledge provided by the GE_{go} corpus also distinctly helps average cross-domain generalizability performance when Offensive Reddit (↑5.7%) and Kumar (↑3.3%) are used as the training sets. The individual cross-domain generalizbility enhancement is most pronounced with the W&H → Razavi (↑15.4%), W&H → Kumar (↑13.6%), and W&H → Founta (↑9.1%).

When we adopt the GE_{ek} corpus with fBERT as the base model in our emotion-enriched framework, we observe performance improvement in both in-domain and cross-domain settings. The greatest performance in-domain uplift is 3.5% with the Founta dataset. The average increase in binary F1 ranges from 0.6% to 8.3% in cross-domain settings with the least and best generalizability corresponding to the Razavi and W&H training sets. Strong average cross-domain generalizability enhancement also manifests in experiments where Founta (↑5.9%) and Offensive Reddit (↑4.0%) datasets are the training sets. The greatest generalizability uplifts for individual cross-domain experiments are shown in W&H → Razavi (↑14.6%), W&H → Kumar (↑14.1%), and Founta → Razavi (↑10.3%). Experiments W&H → Razavi and W&H → Kumar are also the top-performance individual cross-domain experiments using the GE_{go} corpus.

5 Analysis and Discussions

In this section, we study the factors in our framework that affect cross-domain generalizability based on the experimental results.

5.1 Emotion Corpora and Cross-Domain HS Generalizability

Observation 1: *Adopting emotion corpora with fewer categorical scopes, such as the GE_{ek}, generally results in more consistent cross-domain HS generalizability improvement.*

To analyze the relationship between the categorical scope of emotion corpus and cross-domain generalization, we visualize the average performance of the cross-domain generalization of two variants of our emotion-enriched models eliminating the distinction of the employed base models. We take the average performance of the cross-domain generalizability of BERT-based and fBERT-based baselines and emotion-enriched models separated by the use of different emotion corpora. The average performance in binary F1 of the baseline, GE_{go}-based model, and GE_{ek}-based model are shown in Figure 2a. Figure 2b shows the percentage of cross-domain improvement relative to the baseline average.

We observe that the emotion-enriched model

where the emotion knowledge is introduced by the GE_{ek} exhibits better performance in all cases except the experiment where the Offensive Reddit dataset is used as the training set. The highest generalizability improvement by an increase in percentage employed W&H as the training set, resulting in an improvement of 7.4%. We note that the W&H dataset contains relatively the shortest average sentence length, and the HS samples show a direct HS style. These characteristics are also exhibited in experiments when the Founta dataset is employed as the training set, which shows the second highest generalizability improvement by percentage (↑5.1%). Thus, the emotion knowledge supplied by the GE_{ek} corpus is generally the better choice for improving cross-domain generalizability when a model is trained with short, explicit HS and is expected to generalize to HS that could be longer in length where the HS style might also be more implicit (e.g., Razavi, Kumar). For the case where the Offensive Reddit dataset is employed as the training set, we find that the HS samples in this dataset gear toward explicit sexism, which is not perceived in other datasets. In this case, the emotion knowledge supplied by the 28-class emotion corpus GE_{go} helps to mitigate the semantic variance across contrasting HS topics more than the six-class GE_{ek} corpus.

5.2 Domain Adaptability of Base Models and Cross-Domain HS Generalizability

Observation 2: *The strength cross-generalizability enhancement is more pronounced with our emotion-enriched model when adopting a base model that is not adapted to the HS domain (e.g., BERT). However, adopting a base model that is adapted to the HS domain (e.g., fBERT) using our framework generally results in the highest cross-domain performance.*

To analyze the effect of the adopted base models' domain adaptability on cross-domain generalization, we visualize our framework's performance on cross-domain generalization using BERT-based and fBERT-based models eliminating the distinction of emotion corpora. We take the average performance of the cross-domain generalizability average of the variant of our model that uses the GE_{go} corpus and the variant of our model that uses the GE_{ek} for each base model. The average performance in binary F1 for the BERT-based and fBERT-based model are shown in Figure 3a. Figure 3b shows the percentage of cross-domain improvement relative to the baseline average.

From Figure 3, we note that adopting a non-HS domain-adapted model as the base model like BERT with our framework results in the greatest percentage of generalizability improvement in most cases. For the two cases where BERT-based emotion-enriched models show a relatively weaker generalizability uplift than fBERT-based models, we note that the training sets, Founta and W&H datasets, are also the only two datasets that are sourced from Twitter.

Figure 3a supports that adopting a base model like fBERT that is adapted to the HS domain leads to higher performance despite the effect of adding emotion knowledge in uplifting cross-domain generalizability might not be as strong as adopting a non-HS domain adapted base model. For experiments where training sets are Kaggle, Kumar, Razavi, and W&H, our emotion-enriched model that utilizes fBERT, which is pre-trained on a dataset that is in the HS domain, shows the best performance.

From Figure 3b, the most distinct improvement is in cross-domain settings with Kumar (↑7.6%), Offensive Reddit (↑5.7%), and W&H (↑5.6%) datasets as the training sets. As mentioned, the Offensive Reddit dataset exhibits a topical contrast to other datasets as its HS has a sexism focus. Hence, adopting a general base model with emotional knowledge helps to reduce the semantic variance across contrasting HS topics. We observe that the Kumar and W&H datasets both exhibit relatively indirect styles of HS. This suggests that incorporating emotion knowledge helps to bridge the gap in allowing an HS model trained with a general non-HS domain adopted model on implicit HS to generalize on HS that are relatively more direct (e.g. Founta, Kaggle).

6 Conclusion

In this work, we investigated cross-domain HS generalizability integrating emotion analysis. We presented a multitask HS generalizability framework that utilizes emotion knowledge to enhance cross-domain HS generalizability. We employed the 28-class GoEmotions corpus (Demszky et al., 2020) and the traditional six-class Ekman corpus (Ekman, 1971) to examine their effects on improving cross-domain HS generalizability. We found that incorporating emotion knowledge using the Ekman corpus leads to more consistent generalizability performance. We also inspected the role of

HS domain adaptiveness in base models on cross-domain HS generalizability and noted that the introduction of emotion knowledge has a relatively stronger strength in bridging the cross-domain generalization gap of pre-trained models that are not adapted to the HS domain. Results support that our emotion-enriched models outperform baselines in all cross-domain settings.

Limitations

We acknowledge limitations in preserving the conceptual granularity exhibited in public HS datasets by adopting their varied categorical labels (i.e. toxic, abusive, sexism) in a binary form. Furthermore, the analyses presented in this work are based on the chosen datasets corresponding to their domain(s) only. Therefore, conclusions drawn from the limited quantity of datasets from restricted domains are not intended to be comprehensive. We also noted that more potential insights regarding HS generalizability may be drawn by comparing the results from evaluations against more state-of-the-art baselines with varied domain adaptiveness to specific aspects of HS (i.e. implicitness, sarcasm). This is left to future works.

Acknowledgements

We thank all reviewers for their constructive feedbacks. This work is supported by in part by NSF 1946391 "RII Track-1: Data Analytics that are Robust and Trusted (DART): From Smart Curation to Socially Aware Decision Making."

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Federico Bianchi, Stefanie HIlls, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2022. "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8093–8099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia. Association for Computational Linguistics.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 24(4):393–414.

Lisa Hilte, Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2023. Who are the haters? a corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6:986890.

Tal Ilan and Dan Vilenchik. 2022. Harald: Augmenting hate speech data sets with real data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2241–2248, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yifan Liu, Bohan Zhuang, Chunhua Shen, Hao Chen, and Wei Yin. 2019. Training compact neural networks via auxiliary overparameterization. *CoRR*, abs/1909.02214.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Hate speech and offensive language detection using an emotion-aware shared encoder. In *IEEE International Conference on Communications (ICC)*, Roma, Italy, France.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Julian Peller. 2014. Toxic comment classification challenge. *"https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge"*.

Flor Miriam Plaza-del-Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. *arXiv e-prints*, page arXiv:2109.10255.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23*, pages 16–27. Springer.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2019. Attending the emotions to detect online abusive language. *arXiv preprint arXiv:1909.03100*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong

Kong, China. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Washam. 2019. Subreddit-classification. *https://github.com/jwasham12/Subreddit-Classification*.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.