Closest Positive Cluster Loss: Improving the Generalization of Implicit Hate Speech Classifiers Across Social Media Datasets

Saad Almohaimeed, Saleh Almohaimeed, Damla Turgut and Ladislau Bölöni
Dept. of Computer Science, University of Central Florida, Orlando FL 32816

***Abstract*—Flagging hate speech on social media messages has important societal benefits. While large language models have become increasingly able to identify hate speech with high accuracy, they come with a significant computational cost. Thus, there is a need for simpler models that detect hateful or abusive content by classifying an embedding of the text. Such models perform very well for explicitly abusive content, but struggle with the classification of implicit hate. Furthermore, it has been found that the performance decreases significantly in cross-dataset experiments. In this paper, we propose Closest Positive Cluster (CPC) an auxiliary loss that increases the generalizability of embedding-based explicit and implicit hate classifiers in cross-dataset scenarios. Through experiments spanning ten different hate speech datasets, we found that the CPC loss increased the model performance by 0.17 - 7.4% when added to the binary cross-entropy loss during training. The experiments also investigated whether models trained on specific hate speech datasets generalize better to other datasets.**

I. Introduction

There is a consensus that hate speech is a major problem on social media platforms worldwide. Hateful messages are created with an explicit intent to harm, demean or threaten a certain population. Following the viral nature of social media platforms, these messages are rebroadcasted, replied to and quoted by other users. At high density, they can make entire platforms toxic to certain populations, compelling people to retreat into their own communities, leading to a fragmentation of the social discourse, and further political polarization.

It is thus important for all stakeholders in social media to identify potentially hateful messages. Compared to the state of the art of even ten years ago, when lexicon-based approaches predominated, current generation large language models can identify hateful content in even subtle forms with a very high accuracy. Such LLMs, however, with billions of parameters, achieve this with a high computational cost. It is not feasible to pass the entire message traffic of the internet through major LLMs for hate identification.

A more practical choice, pursued by many researchers, is to use a significantly smaller language model (such as $\text{BERT}_{\text{base}}$ [1]) to calculate an embedding that can be classified using a multilayer perceptron or similar architecture. Having only 110M parameters, such classifiers can be run for inference on edge hardware or even mobile devices. Another possibility is to quickly retrain a model with a small dataset of new message samples to account for the changing nature of hate speech. For instance, such a model can be retrained in approximately five minutes on a single NVIDIA Tesla V100 card.

As expected, the performance of embedding-based hate speech classifiers is lower in the case of implicit or subtle hate speech. Furthermore, a significant drop in performance was found when a classifier trained on a specific hate speech dataset was tested on a different dataset [2]. Given the architecture of the models, we conjecture that a possible reason for this is the fact that the embedding does not capture the various language constructs coding for hate. Starting from this hypothesis, several previous projects proposed to improve embedding-based hate speech classifiers by adjusting the embedding to better account for implicit hate speech.

In this paper, we describe a contribution in this direction, by defining a novel auxiliary loss, the Closest Positive Cluster (CPC) loss. Partially inspired by contrastive loss functions, the CPC loss is designed to be combined with the binary cross-entropy (BCE) loss in the training of the classifier.

The main contributions of this paper are as follows:

- Introduce the Closest Positive Cluster auxiliary loss for embedding-based hate speech classifiers.
- Validate the benefits of the CPC loss by training classifiers both in the BCE and BCE+CPC configurations for ten different hate speech datasets, and evaluate the configurations in cross-dataset settings.
- Study the relative generalization performance of different hate speech datasets when used in the training of cross-dataset classifiers.

II. Related Work

The impact of hateful speech on social media has stimulated significant interest from the research community. Researchers have recently examined various aspects of hate speech, including offensive language, cyberbullying, abusive language, and toxic content [3], [4]. Furthermore, several specialized focus areas have also emerged, with particular attention to distinct types of hate speech such as sexism and racism [5], hate speech within sports [6], [7] and political hate speech [8]. Moreover, efforts have expanded beyond just detecting hate speech, such as studies focused on the identification of hate within specific targeted groups [9], [10], [11], the development of explainable models for hate speech detection [12], and the challenges associated with cross-lingual and cross-domain hate speech detection [10], [13], [14].

The focus of this paper is on classifying *implicit hate speech* where the hateful message is conveyed through implications, allusions, use of irony, and other verbal techniques. Implicit hate speech is difficult to detect. Lexicon-based approaches perform poorly, as implicit hate speech often does not contain explicitly offensive words. Furthermore, it represents a challenge even for embedding-based approaches, as an implicitly hateful statement might be superficially formulated as a positive statement.

M. ElSherief et al. [15] was one of the first studies to investigate the classification of implicit hate speech. The paper introduces the IHC dataset of 21k tweets that contains a wide variety of examples of implicit hate speech. To achieve classification, the authors investigated a wide range of candidate technologies, including data augmentation, knowledge graphs, and multi-label classification with a variety of ML models. The experiments found that knowledge graphs were ineffective in detecting implicit hate, with the BERT-based models being the most effective among the investigated technologies.

Caselli et al. [16] introduced HateBERT, a variant of the BERT model retrained on RAL-E, a specialized dataset with a frequent occurrence of hate speech. This dataset was created from about 1.5M anonymized posts to Reddit communities banned for promoting offensive, abusive, or hateful content. HateBERT has been shown to improve the performance of hate speech classifiers when substituted for the original BERT model.

Han and Tsvetkov [17] proposed a technique to strengthen existing implicit hate classifiers without a large annotated dataset. They used probing examples from the SBIC dataset and applied tracking methods (e.g. gradient product, influence functions and training loss) to track the influential samples that led to misclassification and re-annotate them. The gradient product method was found to best enhance the model's performance when applied to implicit hate classifiers.

Ocampo et al. [18] performed an in-depth analysis of implicit and subtle hate speech, reannotating messages from seven datasets of hate speech to further categorize them as explicit or implicit hate speech as well as whether the hate is expressed through subtle means or not. The reannotations also considered attributes such as irony or exaggeration. The experiments compared the original BERT models with variants such as HateBERT and DeBERTa, and the performance of various data augmentation approaches and their combinations.

Almohaimeed et al. [19] explored transfer learning and lexicon-based approaches to enhance the detection of implicit hate speech. The experiments were performed on combinations several datasets. The first set of datasets were of "flagged speech" [3], [9], [12] which consider general labels such as hate and offensive, as well as datasets focused on sexism and racism [5]. These were complemented by datasets such as [4], [10] which provide more nuanced labels such as aggressive, cyberbullying, disrespectful, and fearful speech.

The research most closely related to the approach proposed in this paper aims to improve the performance of embedding-based classifiers by modifying the training / fine-tuning process of the embedding and classification layers.

Kim, Park and Han [2] start from the observation that the performance of implicit hate speech classifiers built on models such as BERT and HateBERT drops significantly when tested on a different dataset. To improve the cross-dataset generalization, the authors propose a contrastive learning approach using two variations. In the AugCon variation, the positive pairs are lexically augmented variations of the original implicit hate speech text. In the ImpCon variation, the positive pairs are the original text and a human annotated implication, where different implicit hate speech expressions are assumed to express the same implication. The experimental results show that the ImpCon model significantly improves the performance in cross-dataset testing.

Ocampo, Cabrio and Villata [20] propose an approach to bring closer the encodings of the explicit and implicit hate speech sentences with a similar target. To achieve this, the BERT and HateBERT encoders were fine-tuned with contrastive learning using as positive examples pairs of implicit and explicit hate speech and as negative examples instances of hate speech and non-hate speech sentences. The new models were found to improve the classification in borderline cases.

III. Closest Positive Cluster Loss

Let us denote with $\mathcal{D} = \{t_1, \ldots, t_n\}$ a dataset of n text messages, that we will consider for classification into hateful and non-hateful speech. Let us consider $\mathcal{E} = \{e_1, \ldots, e_n\}$ the embedding vectors corresponding to these messages. Let be $\mathcal{E}^+ = \{e_1^+ \ldots e_m^+\}$ the subset of m embeddings that correspond to the positive (hate speech) samples.

We will develop a custom loss function starting from the conjecture that the concept of hate speech, especially in its implicit form, is *inherently multimodal*. Instead of aiming to bring all the instances of implict hate speech to the same area of the embedding space, we will allow multiple clusters of hate speech to exist in the embedding.

We start by clustering the positive samples $\mathcal{E}^+$ into k clusters using the k-means algorithm. Experimentally we found that $K = \sqrt{|\mathcal{E}^+|}$ provides the best performance. For each cluster j, we will denote with $C^j = \{e_1^{j+} \ldots\}$ the embeddings in the cluster (which, due to the construction, will all be positive). We define the center of the cluster c_j and its radius r_j by:

$$c_j = \frac{1}{|C^j|} \sum_{e^+ \in C^j} e^+ \tag{1}$$

$$r_j = \max_{e^+ \in C^j} \|e^+, c_j\|_2 \tag{2}$$

where $\|\cdot,\cdot\|_2$ is the euclidean distance between the embedding vectors. To capture the intuition that embeddings within the radius are part of the cluster, while embeddings that are several radii away are not, we define the *normalized distance*:

$$\|e, c_j\|_n = \frac{\|e, c_j\|_2}{r_j} \tag{3}$$

In the remainder, we will say that an embedding e is *inside* the cluster centered on c if $\|e, c\|_n < 1$, and is *outside* the cluster

if $\|e, c\|_n >= 1$. If we denote the index p of the cluster whose center is closest:

$$p = \underset{i}{\operatorname{argmin}} \left(\|e, c_i\|_2 \right) \tag{4}$$

we define the *normalized distance from the closest cluster* as:

$$\|e\|_{cc} = \frac{\|e, c_p\|_2}{r_p} \tag{5}$$

The various components of these definitions are illustrated in Figure 1. Starting from these definitions we are introducing a custom loss function we call the Closest Positive Cluster (CPC) loss.

Inspired by the contrastive learning loss function [21], which encourages the model to generate similar embeddings for similar samples while producing distinct embeddings for different samples, our proposed function follows a similar principle. Specifically, our function aims to encourage the model to generate similar embeddings for similar samples by utilizing clusters of positive samples. Additionally, it enhances the flexibility of embeddings by allowing negative sample embeddings to be positioned closer to positive clusters but outside the clusters, ensuring coverage of neutral samples that are semantically closer to hate samples. So, the intuition behind the CPC loss is as follows:

- For positive samples that are inside at least one positive cluster, and negative samples that are outside all positive clusters, the loss is zero.
- For a positive sample that is outside all positive clusters, we aim to bring the node toward the closest positive cluster.
- For a negative sample that is inside a positive cluster, we aim to push the sample away from the cluster center.

Let us now consider a text t and a parameterized language model that generates the embedding $e = M(t; \theta)$. The ground truth label for this text is $y = 1$ for positive samples and 0 for negatives, while the output of the classifier is $\hat{y} = P(e; \phi)$. We define the CPC loss as follows:

$$\mathcal{L}_{CPC}(t, y, \theta, \phi) = y \cdot \max\left(\|e\|_{cc} - 1, 0\right) + (1 - y) \cdot \max\left(1 - \|e\|_{cc}, 0\right) \tag{6}$$

Note that the CPC loss does not directly improve the prediction accuracy, that is the match between the correct label y and the classifier output $\hat{y}$. Thus, for our complete system we use a loss that combines the CPC loss with the well known binary cross-entropy (BCE) classification loss:

$$\mathcal{L}(t, y, \theta, \phi) = \mathcal{L}_{CPC}(t, y, \theta, \phi) + \mathcal{L}_{BCE}(t, y, \theta, \phi) \tag{7}$$

The overall training pipeline is described in Figure 2.

IV. Experiment Settings

For the experiments in this paper, we use seven datasets of explicit hate speech and three datasets of implicit hate speech. Table I describes the list of datasets, the papers where they were first described, and the number of positive and negative samples in the dataset. For the convenience of reference, when the dataset was not given a specific name, we refer to it by the first author of the introducing paper. Different datasets often use slightly different labels and terminology to refer to hate speech messages. To avoid this source of confusion, we use only two classes for explicit hate speech: normal and flagged, where flagged speech includes hate speech, offensive speech, or any other suspicious speech that might represent abusive language. We also use only two classes for labeling implicit hate: non implicit-hate (NIH) and implicit hate (IH). We followed the label unification done in [19] for this purpose.

TABLE I: The explicit and implicit hate speech datasets used in our experiments, along with the number of positive and negative samples in each, are presented. We consider only English-language, as some datasets contain multilingual samples.

(a) Flagged (or Explicit) Speech Datasets

Dataset Name	Normal	Flagged
Waseem [5]	10423	6484
Davidson [3]	2641	22142
Founta [4]	38952	7030
OLID [9]	8382	4858
Ousidhoum [10]	371	5276
hateXplain [12]	4606	15542
THOS [11]	5826	2456

(b) Implicit Hate Speech Datasets

Dataset Name	NIH	IH
THOS_IH [11]	6538	1744
IHC [15]	14380	7100
OLID_IH [22]	11742	1498

As seen in Table I, different datasets have different sizes and a different mix of negative and positive samples. While, obviously, everything else being equal, a larger training dataset is preferable, our objective is to investigate the proposed training methodology in the context of the cross-dataset nature of the testing. What we are primarily interested in is the fact that the different datasets might capture different types of hate speech. To allow us to study this, we first equalize the size of the datasets in the following way. When a dataset is used for training, from each dataset we randomly sample 200 positive and 200 negative samples for training, and 50+50 samples for validation. When a dataset is used for testing, we randomly choose 500+500 samples. Due to the limited availability of either positive or negative annotated data in the testing datasets, we find that 1,000 samples are sufficient for all testing datasets except for Ousidhoum, where the number of testing samples is reduced to 742 due to the scarcity of negative samples we have from this study.

Furthermore, in cross-dataset testing settings, it is common practice to use five random seeds, as applied in [2], [15], [17], [20]. Based on our observations, averaging the results from three random seeds is sufficient when a large number of testing samples (e.g., 5,000 samples) is available. However, for testing with 1,000 samples, we observed that using four to

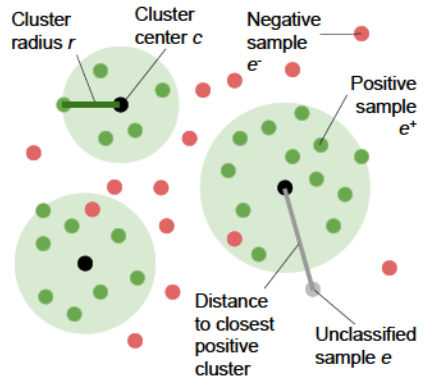


Fig. 1: A visual illustration of the terms contributing to the definition of the CPC Loss

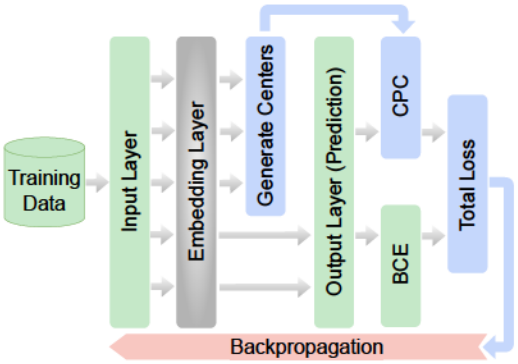


Fig. 2: The training pipeline of hate speech classifier using a combination of BCE and CPC loss.

five random seeds produces more consistent results. Based on this observation and the common practice in previous studies, all the results presented in the following section are the result of averaging 5 random seeds for every experiment.

For all the experiments we report, we use the $\text{BERT}_{\text{base}}$ uncased pre-trained model. We experimented with learning rates of 2e-5, 5e-6 and 2e-6, 1-30 epochs and the batch size 4 and 8, reporting the results for the best performing combination.

V. Results

A. Classification accuracy improvement by the CPC auxiliary loss

Our first set of experiments investigates whether the addition of the CPC loss improves the accuracy of the learned classifier. We are interested both on the on-dataset and the cross-dataset settings, with the latter being known to negatively affect the performance. Tables II and III contain these results for the explicit and implicit hate speech datasets, respectively. We found that in 69% of the experiments, the CPC+BCE combination improves the F1 score, 10.3% it does not affect it, while in 20.7% of the cases it performs worse than the BCE loss alone. The improvement, when present, ranges between 0.17 and 7.4%.

B. Datasets As Training

One aspect of cross-dataset testing of hate speech classifiers trained within the same type of datasets (e.g. training on explicit hate dataset and testing on other explicit hate datasets) is that certain datasets might be better suited as training data, even when tested on other datasets. This might possibly be due to their covering a wider range of hate speech constructs, or at least expressing them in a way that is easier for the embeddings to capture. An implication of this might also be that the relative performance of a dataset in training might be affected by an auxiliary loss such as CPC. Fig. 3a and 3b show the results of experiments when one dataset was used as training data which was then evaluated (a) on test data drawn from the same dataset and (b) on test data drawn from other datasets. As expected, the performance is lower in the cross-dataset settings. The figures show this as the penalty in the macro-averaged F1 score, which is normally a negative percentage. The closer this number is to zero, the better the dataset is as a training set.

Fig 3a considers the explicit hate speech datasets. We find that Davidson, Founta, OLID, and THOS have the lowest penalty. Another interesting observation is that for most datasets, the use of the BCE+CPC combination lowers the cross-dataset penalty. Fig 3b considers the implicit hate speech datasets. Overall, we found the cross dataset penalty to be greater for implicit hate compared to the explicit datasets (verifying, once again, the observation made in [2]). We find that in all cases, the use of the CPC auxiliary loss allows for a lower penalty.

C. Datasets as Testing

The converse of the performance of a hate speech dataset as training data, is the performance of the classifiers when the dataset is used for testing classifiers trained on other datasets within the same type. There can be a significant performance variation among datasets: for instance, some datasets might naturally have more borderline samples, contain aspects of hate that are less covered by other datasets, or make different cutoffs in labeling ambiguous cases. Fig. 4a and 4b show the results as a penalty in the macro-averaged F1 score, which are, as in the previous case, negative percentages.

Fig 4a shows the results for explicit hate datasets. An interesting observation is that Davidson, the dataset that obtained the best performance as a training dataset for other datasets, had the highest penalty when tested with classifiers trained on other data. For the implicit hate datasets in Fig 4b we see a similar observation: the THOS_IH dataset, which was comparatively the best for training, gives the lowest results for testing. We note, however, that this inverse relationship does not apply for every dataset. We also note that in seven out of ten cases, the use of the CPC auxiliary loss reduces the penalty of cross-dataset testing.

D. Training With Explicit Datasets and Testing on Implicit Dataset

The challenges of collecting and labeling implicit hate speech datasets make it relevant to ask the question whether classifiers

TABLE II: Comparison of the use of the BCE loss only versus the BCE+CPC losses for cross-dataset training for explicit hate speech detection. The values presented are the macro-averaged F1 score. The better performing approach is in **bold**. Test results for the case where the model was trained and tested with the same dataset are between parenthesis.

↓ Train Dataset	Test dataset→	Waseem	Davidson	Founta	OLID	Ousidhoum	hateXplain	THOS
Waseem	BCE	(75.3)	74.2	79.7	70.7	65.9	64.7	78.6
	BCE+CPC	**(75.4)**	**75.1**	**80.7**	70.7	**67.2**	**69.0**	**80.5**
Davidson	BCE	62.5	(92.9)	**79.6**	67.7	73.8	73.6	91.6
	BCE+CPC	62.5	**(93.4)**	79.4	**68.1**	**74.9**	**74.5**	**93.2**
Founta	BCE	71.9	76.5	**(85.2)**	75.1	69.9	**71.5**	87.9
	BCE+CPC	**73.7**	**77.4**	(85.1)	**76.6**	**71.4**	69.5	**90.1**
OLID	BCE	68.8	77.1	83.7	(75.8)	71.6	71.0	90.2
	BCE+CPC	**71.1**	**78.5**	**84.9**	**(78.1)**	**72.6**	71.0	**92.2**
Ousidhoum	BCE	54.7	**82.9**	75.5	**64.5**	**(78.0)**	**73.5**	83.2
	BCE+CPC	**55.6**	79.7	75.5	63.8	(77.2)	73.4	**83.5**
hateXplain	BCE	60.4	**78.4**	77.3	65.3	**74.4**	**(76.9)**	83.2
	BCE+CPC	**62.2**	78.0	**79.1**	**68.6**	73.8	(76.7)	**85.9**
THOS	BCE	60.7	79.9	81.0	71.0	72.1	74.1	(93.6)
	BCE+CPC	**63.1**	**80.9**	**82.0**	**72.0**	**73.9**	**74.3**	(93.6)

TABLE III: Comparison of the use of the BCE loss only versus the BCE+CPC losses for cross-dataset training for implicit hate speech detection. The values presented are the macro-averaged F1 score. The better performing approach is in **bold**. Test results for the case where the model was trained and tested with the same dataset are between parenthesis.

↓ Train Dataset	Test Dataset →	THOS_IH	IHC	OLID_IH
THOS_IH	BCE	(74.4)	52.4	58.4
	BCE+CPC	**(74.8)**	**54.4**	**59.8**
IHC	BCE	50.8	**(64.4)**	51.4
	BCE+CPC	50.8	(63.0)	**52.3**
OLID_IH	BCE	59.1	54.0	**(60.8)**
	BCE+CPC	**59.5**	**58.0**	(59.6)

TABLE IV: Macro-averaged F1 score for a classifier trained on explicit datasets and tested on implicit datasets.

↓ Train Dataset / Test Dataset →	THOS_IH	IHC	OLID_IH
Davidson	45.2	52.2	48.0
Founta	**52.2**	**59.6**	**59.0**

trained on explicit hate speech datasets can classify implicit hate speech. The reason why this might be possible in principle is both the fact that some explicit datasets also contain samples without explicitly offensive words. In addition, the embedding layer might map examples of implicit hate close to explicit hate in the embedding space.

To investigate this, we considered Founta, Davidson and THOS, the datasets that performed best as training data. From this list, we discarded THOS to avoid model bias, as THOS_IH has the same rows as THOS with different annotations. The results of these experiments are shown in Table IV. We find that the results obtained with Founta are significantly better than those with Davidson when tested with all three implicit hate datasets.

VI. Conclusion

In this paper, we described CPC, a novel auxiliary loss function to enhance the performance of an embedding-based binary hate speech classifier. Through extensive experiments across ten datasets, we demonstrated that using CPC in combination with the BCE loss can lead to an improvement of up to 7.4% in model performance in cross-dataset settings.

To better understand the performance of hate speech classifiers, both with the BCE and CPC+BCE losses, we carefully studied the composition of ten explicit and implicit hate speech datasets, and the impact this composition and the presence of explicitly offensive words have on their performance as a training dataset and difficulty to classify when used as a test dataset. Our key finding was that datasets that have a minimal amount of mislabeling and a diverse range of explicit and implicit positive samples (for explicit datasets) or a diverse range of normal and offensive negative samples (for implicit hate datasets) offer enhanced generalizability.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

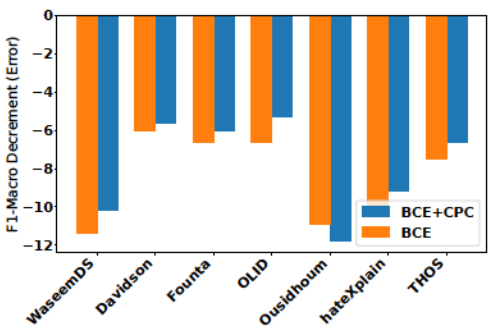
[2] Y. Kim, S. Park, and Y.-S. Han, "Generalizable implicit hate speech detection using contrastive learning," in *Proc. of the 29th International Conf. on Computational Linguistics (COLING-2022)*, 2022, pp. 6667–6679.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

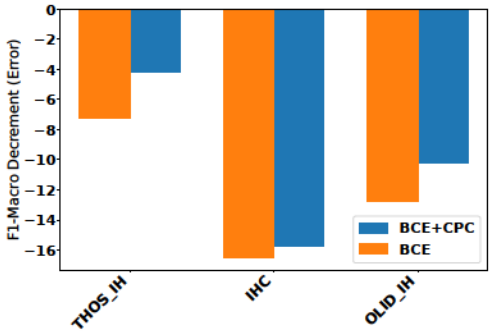
[4] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, vol. 12, no. 1, 2018.

[5] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proc. of the NAACL Student Research Workshop*, 2016, pp. 88–93.

[6] S. Vujičić Stanković and M. Mladenović, "An approach to automatic classification of hate speech in sports domain on social media," *Journal of Big Data*, vol. 10, no. 1, pp. 1–16, 2023.

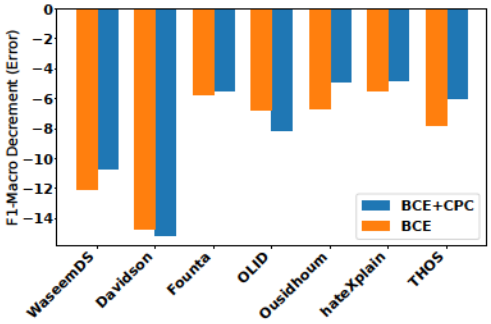


(a) Explicit as Training Datasets

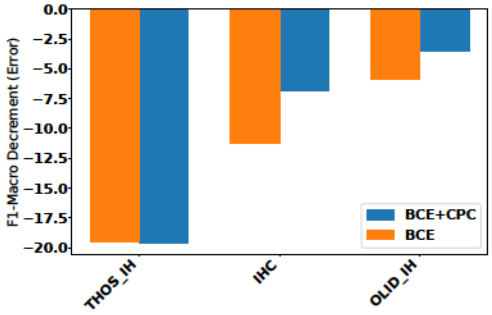


(b) Implicit as Training Datasets

Fig. 3: Performance of different datasets as training data. The data shows the decrease in the macro-averaged F1 score when the classifier is trained on the specific dataset and tested across datasets compared to testing on the same dataset.



(a) Explicit Hate Datasets



(b) Implicit Hate Datasets

Fig. 4: Penalty of cross-dataset training versus training on the same dataset. The data shows the decrease in the macro-averaged F1 score for testing on the given dataset when the classifier was trained on different datasets compared to the same dataset.

[7] E. K. Klutse, S. Nuamah-Amoabeng, H. Lyu, and J. Luo, "Dismantling hate: Understanding hate speech trends against nba athletes," in *Proc. of Int. Conf. on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2023, pp. 74–84.

[8] P. Agarwal, O. Hawkins, M. Amaxopoulou, N. Dempsey, N. Sastry, and E. Wood, "Hate speech in political discourse: A case study of UK MPs on Twitter," in *Proc. of the 32nd ACM Conf. on hypertext and social media*, 2021, pp. 5–16.

[9] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. of NAACL-2019*, 2019, pp. 1415–1420.

[10] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP-2019)*, 2019, pp. 4675–4684.

[11] S. Almohaimeed, S. Almohaimeed, A. A. Shafin, B. Carbunar, and L. Bölöni, "THOS: A benchmark dataset for targeted hate and offensive speech," in *Proc. of Data-centric Machine Learning Research (DMLR) Workshop at ICML 2023*, 2023.

[12] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[13] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Jul. 2019, pp. 363–370.

[14] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: experimenting with Hindi and English social media," *Information*, vol. 12, no. 1, p. 5, 2020.

[15] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 2021, pp. 345–363.

[16] T. Caselli, V. Basile, J. Mitrovic, and M. Granitzer, "HateBERT: Retraining BERT for abusive language detection in English," in *Proc. of the 5th Workshop on Online Abuse and Harms (WOAH-2021)*, 2021, pp. 17–25.

[17] X. Han and Y. Tsvetkov, "Fortifying toxic speech detectors against veiled toxicity," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP-2020)*, 2020, pp. 7732–7739.

[18] N. Ocampo, E. Sviridova, E. Cabrio, and S. Villata, "An in-depth analysis of implicit and subtle hate speech messages," in *Proc. of the 17th Conf. of the European Chapter of the Association for Computational Linguistics (EACL-2023)*, 2023, pp. 1997–2013.

[19] S. Almohaimeed, S. Almohaimeed, and L. Bölöni, "Transfer learning and lexicon-based approaches for implicit hate speech detection: A comparative study of human and GPT-4 annotation," in *Proc. of the IEEE 18th International Conf. on Semantic Computing (ICSC-2024)*, 2024, pp. 142–147.

[20] N. Ocampo, E. Cabrio, and S. Villata, "Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations," in *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 6626–6637.

[21] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR-2006)*, vol. 2, 2006, pp. 1735–1742.

[22] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in *Proc. of the Twelfth Language Resources and Evaluation Conf.* European Language Resources Association, 2020, pp. 6193–6202.