

# Specializing General-purpose LLM Embeddings for Implicit Hate Speech Detection across Datasets

Vassiliy Cheremetiev\*

EPFL

Lausanne, Switzerland

Idiap Research Institute

Martigny, Switzerland

v.cheremetiev@gmail.com

Quang Long Ho Ngo\*

EPFL

Lausanne, Switzerland

Idiap Research Institute

Martigny, Switzerland

quang.ngo@epfl.ch

Chau Ying Kot

EPFL

Lausanne, Switzerland

chau-ying-kot@hotmail.com

Alina Elena Baia

Idiap Research Institute

Martigny, Switzerland

alina.baia@idiap.ch

Andrea Cavallaro

EPFL

Lausanne, Switzerland

Idiap Research Institute

Martigny, Switzerland

andrea.cavallaro@epfl.ch

## Abstract

Implicit hate speech (IHS) is indirect language that conveys prejudice or hatred through subtle cues, sarcasm or coded terminology. IHS is challenging to detect as it does not include explicit derogatory or inflammatory words. To address this challenge, task-specific pipelines can be complemented with external knowledge or additional information such as context, emotions and sentiment data. In this paper, we show that, by solely fine-tuning recent general-purpose embedding models based on large language models (LLMs), such as Stella, Jasper, NV-Embed and E5, we achieve state-of-the-art performance. Experiments on multiple IHS datasets show up to 1.10 percentage points improvements for in-dataset, and up to 20.35 percentage points improvements in cross-dataset evaluation, in terms of F1-macro score.

**Content warning:** This paper discusses examples of harmful text that may be offensive or upsetting.

## Keywords

implicit hate speech, detection, context, embeddings

## 1 Introduction

Hate speech detection is important to support content moderation in digital platforms, to foster inclusive discourse and to prevent social harm.

Hate speech can be explicit or implicit. Explicit hate speech (EHS) directly targets a protected entity and contains explicit keywords. Hence, early efforts for EHS detection primarily focused on identifying explicitly abusive language through keyword-based approaches [9, 49, 58]. IHS is *"the use of coded or indirect language such as sarcasm, metaphor, and circumlocution to disparage a protected group or individual, or to convey prejudicial and harmful views about them"* [12, 15, 57]. IHS has a nuanced nature and manifests through a diverse range of subtle forms such as stereotypes, humor, and sarcasm [8, 14, 24, 45, 48, 57]. Although IHS may not contain explicit hate words, it propagates prejudice and discrimination, and it is equally harmful as its explicit counterpart [5, 40]. Even humans

may struggle to understand the underlying meaning and intent behind such expressions [18, 48].

Detecting IHS is made difficult by its lexical and semantic similarity to non-hateful content. IHS detection requires a nuanced understanding of implied meaning [35], real-world knowledge related to an event, specific social contexts, and the target.

LLMs capture and represent extensive world knowledge [63], which could be leveraged for hate speech detection. Prior works explored prompting LLMs in scenarios like zero-shot [7, 20, 30, 61, 67], zero-shot with chain-of-thought [61], and few-shot in-context learning [65]. LLMs incorporate safeguards that prevent models from answering or discussing some sensitive topics like hateful content. Moreover, LLMs may exhibit limitations like excessive focus on sensitive groups, thus resulting in wrong classification of benign speech as hate, or extreme confidence score distributions resulting in poor calibration [65]. Overall, these models (e.g., GPT-3.5-Turbo, LLaMa2-7B, Mixtral-8x7b) typically underperform task-specific fine-tuned models [7, 61, 65].

In this paper, we evaluate fusing multiple sources of information to enhance BERT-based classifiers and leverage the ability of LLMs to generate contextual information for IHS detection. Specifically, we explore four fusion strategies to complement content information with contextual and emotion information. We find that while information fusion via feature concatenation provides a slight improvement over content-only BERT-based classifiers, fine-tuning general-purpose LLM-based embeddings (e.g., Stella [64], Jasper [64], NV-Embed [29], E5 [55]) allows us to reach new state-of-the-art performance for IHS detection. In summary, our main contributions are as follows:

- We present a comprehensive comparative evaluation of BERT-based and recent embedding-based classifiers, and show that fusion with LLM-generated context and emotion information can only marginally enhance the performance of a BERT-based classifier. We introduce new state-of-the-art benchmarks in this category of classifiers based on fine-tuning of generalist embedding models.

\*These authors contributed equally to this research.

- We show that specializing embedding-based models significantly improves IHS detection in cross-dataset settings. This approach outperforms current state-of-the-art methods [23, 25, 27, 61] on several IHS datasets up to 1.10 percentage points for in-dataset evaluation and up to 20.35 percentage points for cross-dataset evaluation (F1-macro score). The significant improvement in cross-dataset evaluation is particularly noteworthy for generalization across datasets.

Our approach is significant because it simplifies the detection process and eliminates the need for (explicit) external knowledge. To the best of our knowledge, we are the first to use general-purpose LLM-based embeddings models for IHS detection. The code is available at <https://github.com/idiap/implicit-hsd>.

## 2 Related Work

Early research in hate speech detection primarily focused on identifying explicit abusive language through linguistic features, such as character n-grams [58] or word-centered features (i.e., literal words, part-of-speech tagging, occurrence of words within a word window) [56]. A combination of features such as TF-IDF weighted n-grams, part-of-speech tags, metadata including indicators for elements like hashtags and URLs, and number of characters and words was also used to train classifiers [9, 49]. In [10], the authors explore the combination of lexical and syntactic features with word sentiments and word embeddings. These models rely on phrase structure and fail to capture the complexity and subtlety of the language used in social media. Transformer-based models have improved the quality of classification [39, 47]. Later works [8, 14, 24, 45, 48, 57] have emphasized the nuanced nature and complexity of implicit hate. Progress has been made in this area by focusing on specific types of implicit hate, such as euphemistic hate speech [34], sarcasm detection [1], as well as through multi-task learning [4, 21, 37, 38, 43], external knowledge integration [27, 31, 42, 50, 61] or contrastive learning-based methods [2, 23, 25, 41].

**Multi-task learning.** Classifiers can be trained to detect hate speech jointly with secondary tasks. For example, as hate speech may relate to emotions [13], a secondary task can be emotion classification [37]. Plaza-Del-Arco et al. [43] achieves promising results on binary hate speech detection by combining sentiment and emotion into their features. Awal et al. [4] employs a multitask learning approach to jointly learn hate speech detection with secondary tasks, such as emotion classification and hateful target identification. The authors use a BERT transformer [11] to share knowledge between tasks and Bidirectional Long-Short Term Memory Networks to learn task-specific representation, followed up by a gated fusion mechanism. The authors base their approach on the intuition that datasets from relevant tasks can augment the hate speech data for the primary detection task. The method proposed in [38] leverages emotion recognition as an auxiliary task for both hate speech and offensive language detection, via a shared BERT-based encoder and task-specific classification heads. Similarly, Jafari et al. [21] incorporates sentiment features alongside fine-grained emotion and textual features to improve the detection of IHS compared to single-task methods.

**External knowledge.** Recent research focuses on enhancing hate speech detection by integrating various forms of real-world external knowledge (entity linking [31], knowledge bases [50]). Lin et al. [31] links words appearing in tweets to their Wikipedia description and concatenates them with the original tweet before encoding. Sridhar et al. [50] combine explicit knowledge from knowledge bases with expert knowledge from high-quality annotation and LLM-generated knowledge to improve explanations of stereotypes in toxic speech. Kim et al. [26] and Kim et al. [27] propose methods that utilize external knowledge, such as implications of anchor sentences and synonym substitution or machine-generated statements, respectively, to improve IHS detection using contrastive learning. In [61], the authors incorporate explanations generated using chain-of-thought to better discern between hate and not hate and to improve generalization to unseen datasets. Pérez et al. [42] also demonstrates that hateful messages directed at certain communities, such as the LGBTI community, may benefit from the addition of context. The authors show that incorporating contextual parent comments and the corresponding news articles can improve the detection of hate speech in responses to posts from media outlets.

**Contrastive learning.** Ahn et al. [2] designed a clustering-based contrastive learning technique that uses shared semantics extracted from the data to learn discriminative representations. Specifically, the model is trained to pull together posts from the same cluster and push apart those from different clusters. This approach eliminates the need for costly human-annotated implications or machine-augmented data. Kim et al. [25] propose a contrastive learning-based approach that leverages hard negative samples to mitigate overfitting and improve generalization without relying on external knowledge. Building on this idea, Jiang et al. [23] use prediction errors to select hard positive samples for contrastive learning to encourage the model to learn more robust representations to the spurious attributes that cause the misclassification.

Ocampo et al. [41] use contrastive learning to bridge the representation gap between explicit and implicit hate speech. The authors build upon the observation that explicit and implicit text representations, when grouped by their target groups, tend to cluster together. The method pushes closer together pairs of implicit and explicit messages sharing the same target group, while pushing apart negative pairs (hate and not hate instances). This leads to more meaningful embedding representations and better separations between not hate and hate instances. Masud et al. [35] proposes to improve IHS detection by aligning the surface form of implicit hate with its implied meaning and increasing inter-cluster separation in the latent space to better distinguish speech categories.

## 3 Models

### 3.1 Enhancing BERT-based classifiers

BERT [11] and its variants such as RoBERTa [32] have been extensively used for text classification [3]. Hate speech detection works [2, 23, 25–27] predominantly use models such as BERT, RoBERTa, and T5 [46]. Table 1 shows a summary of the backbone architectures used by the most recent related works on IHS. We enhance the BERT model by incorporating tweet-level emotion information and tweet-driven contextual information via dedicated

**Table 1: Backbone models used for IHS detection. Multiple models indicate variations in the original work.**

Backbone model	Related work
BERT	ImpCon [26], LAHN [25], SharedCon [2], CCL [23], ConPrompt [27], MTL [38], AngryBERT [4], FiADD [35], EHSor [37] Contrastive BERT [41]
RoBERTa	ImpCon [26], LAHN [25]
HateBERT	ImpCon [26], CCL [23], FiADD [35], Contrastive HateBERT [41]
mBERT	MTL [38]

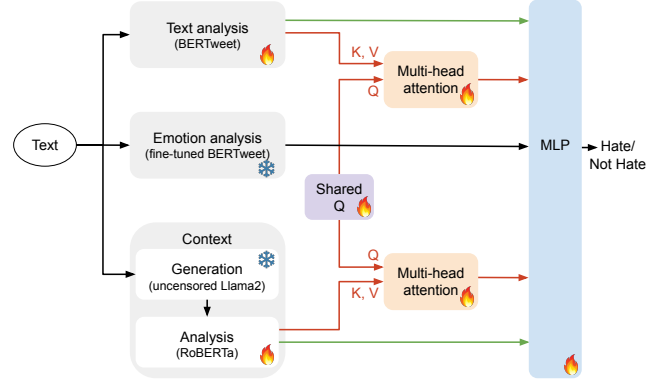
modules. Our BERT-based classifiers consists of three main components, namely text analysis, emotion analysis, and context generation (see Figure 1).

**Feature extraction.** The *text analysis* module uses a fine-tuned BERT to extract the content of the tweet and represent it into an embedding vector. The *emotion analysis* module infers with a fine-tuned BERTweet [44] a vector of probabilities across the following classes: fear, disgust, surprise, anger, sadness, joy, or other. Using a vector of probabilities instead of a single class allows the model to capture the complexity of the emotion. Understanding IHS relies heavily on contextual nuances. Capturing relevant context is made challenging by the short text length (tweets). Our *context* module leverages uncensored Llama2<sup>1</sup> to generate the associated context, avoiding safeguards that might prevent processing and generation of certain content. We prompt the LLM to produce a neutral and factual context, which may include historical background or descriptions of stereotypes concerning the target of the text:

Prompt: *As an educational assistant, your task is to provide neutral and objective analysis of the provided tweet, without any personal biases. Offer short and concise information, context, and concepts to understand the content of the tweet without bias. The tweet may originate from different extremist groups, including White Nationalist, Neo-Nazi, Anti-Immigrant, Anti-Muslim, Anti-LGBTQ, KKK as well as non-extremist sources. The tweet could contain sarcasm, stereotypes, satire, metaphor, irony, or misinformation. Remember to avoid injecting personal opinions or interpretations into your analysis. Your aim is to provide a neutral understanding of the tweet’s content within a maximum of 150 words.*

The final prompt is [Prompt. "Tweet to analyze: ", <Original tweet>]. We explicitly ask for an objective and neutral analysis to try to avoid bias from the data Llama2 was trained on. We also give a context about the dataset that is used so the LLM has a starting point (see Appendix A for examples of generated context). The generated context is then used by RoBERTa to extract features.

**Feature fusion.** We explore four feature fusion approaches, namely concatenation, adaptive fusion, mixture of experts, and shared learnable query. With *concatenation*, we classify with a two-layer perceptron (MLP) the outputs of the three modules stringed together. The first layer of the MLP has the same size as the concatenated embeddings (1543), whereas the second layer contains 2 nodes for the binary classes.



**Figure 1: Overview of our BERT models.** In gray, the three main components: tweet, emotion, and context module. In orange/purple, the added components for the shared learnable query architecture. Green/red arrows show the information flow for the concatenation/shared query fusion, respectively. Emotion features are directly fed to the MLP for both strategies. Q, K, V represent the query, key and value.

With *adaptive fusion*, we learn the parameters  $\alpha_{tweet}$ ,  $\alpha_{context}$ , and  $\alpha_{emotion}$  that determine the scaling of each feature component. In order to maintain reasonable magnitude in the inputs, we constrain to  $[-1, 1]$  the learnable parameters with a sigmoid. With a simple *mixture of experts*, given a short text input, we utilize a simple MLP followed by a softmax layer to generate three adjustable feature scaling factors:  $\alpha_{tweet}$ ,  $\alpha_{context}$ ,  $\alpha_{emotion}$ . The key distinction from adaptive fusion lies in the ability to tailor these scaling parameters specifically for each input, whereas adaptive fusion employs a fixed set of scaling parameters across all samples in the test dataset. Finally, for the *shared learnable query*, we use a multi-head attention with a shared learnable query, where keys and values are derived from both content and context embeddings. The query is a learnable parameter that is the same for both the content and context. The outputs of the multi-head attention blocks are then concatenated along with the emotion vector and fed to the classifier.

### 3.2 Specializing generalist embeddings

General text embedding models, such as Stella [64], E5 [55], NV-Embed [29], and Jasper [64] are the result of numerous improvements over BERT [11] and RoBERTa [32]. Several factors contribute to the better performance of newer embedding models compared to BERT. First, the embedding models are trained on a bigger volume of data than BERT, enabling them to capture more diverse linguistic patterns and contextual nuances. Secondly, techniques such as hard-negative, in-batch negative and contrastive learning in general appear to provide better embeddings for classification even without a task specific pipeline for classification. E5 [55] is initialized from XLM-RoBERTa-large [6] and results from curated datasets and contrastive learning with mined hard negatives. NV-Embed [29], a fine-tuned version of Mistral 7B [22], is trained with contrastive learning using in-batch hard negatives and uses a latent attention layer to produce embeddings. Stella [64] is based on mGTE [66] and the general text embedding variant of Qwen2 [60]

<sup>1</sup>[https://huggingface.co/georgesung/llama2\\_7b\\_chat\\_uncensored](https://huggingface.co/georgesung/llama2_7b_chat_uncensored)



**Figure 2: Overview of the embedding-based models. Given a task specific instruction, the generalist embeddings models are fine-tuned on the IHS datasets.**

where a final training involves matryoshka representation learning (MRL) [28] which makes it performant at different embedding sizes. Jasper [64] uses a distillation of multiple teachers [29, 64] and is augmented with multi-modal capabilities through a final training stage where image-caption pairs are used with SigLIP [52] as the image encoder. These models also come in different sizes, with E5-large at 560 million parameters, Stella at 1.5 billion, Jasper at 2 billion, and NV-Embed at 7 billion.

To remove instruction bias, all models are fine-tuned using the following instruction template: Instruct: classify the following in no hate or hate. \nQuery:. The instruction is prepended to the short text that is being classified and then passed to the general text embedding model. Each model produces embeddings in  $\mathbb{R}^{k \times n}$  whose dimensions depend on their specific implementation and the input length  $k$ . Following the recommendations provided by the model authors<sup>2,3</sup>, we combine these embeddings into a single representation using a normalized sum over the token dimension. NV-Embed uses mean pooling as part of its final layer, we therefore use the output directly. This results in a final embedding vector in  $\mathbb{R}^n$ , which is subsequently fed into the classification module, which consists of a two-layer MLP with a hidden layer of size  $n$  and LeakyReLU activations. The MLP ultimately reduces the dimensionality to 2 for classification (see Figure 2).

To contrast the results of our embeddings-based classifiers, we compare them with linear probing (i.e., only the classification module is optimized) and to recent generative models, such as Llama3-8B [16], Gemma-7B [51], and Qwen3-8B [59]. For these LLMs, we take the average over the last hidden state as our embeddings [54] which are then fed to the same classification module as for the generalist embedding models. We fine-tune the whole pipeline.

## 4 Validation

### 4.1 Datasets

To quantify the performance of the classifiers, we employ four commonly used IHS datasets that cover a variety of contexts and nuances of real-world scenarios. The distribution of labels in each dataset is reported in Table 2.

**Implicit Hate Corpus (IHC)** [12]. This dataset consists of tweets collected between 2015 and 2017 from accounts of US extremist groups, including Black Separatist, White Nationalist, Neo Nazi, Anti-Muslim, Racist Skinhead, Ku Klux Klan, Anti-LGBT and Anti-Immigrant. Most of their speech targets minorities or specific groups of people. The samples are labeled as explicit hate, implicit hate, or not hate. It is important to note the class imbalance in this dataset: 13206 tweets are not hate and 5460 contain implicit

**Table 2: Distribution of labels in the datasets.**

Dataset	# Samples	Hate	Not hate
IHC [12]	18666	5460	13206
Dynahate [53]	41144	22175	18969
SBIC [48]	44781	24048	20733
ToxiGen [18]	9900	3774	6126

hate. Following [25, 26], we only used the implicit hate samples in the dataset as the hate class, meaning that we do not use the explicit hate samples. An example of not hate sample is: "i have no idea what you are talking about. white supremacy = pure evil". An example of implicit hate sample is "#hannahcornelius - why not come home to #europe whites will never be welcome in #southafrica".

**DynaHate** [53]. This dataset is built with an iterative process between a model and human annotators who progressively generate more challenging examples to trick the model (i.e., by flipping labels with minimal changes to the original post). The examples that are successful in tricking the model are then added to the training set. The model used for classification is RoBERTa with a sequence classification head, which is used to evaluate the difficulty of samples. The labeling includes hate/not hate, type of hate (e.g., threat, dehumanization), and target of hate. There are 41,255 entries, with 54% of them labeled as hate.

**SBIC** [48]. This dataset contains social media posts from Reddit and Twitter with implicit social biases, stereotypes, and power dynamics in language. It was annotated by Amazon Mechanical Turk workers. The main labels contain: offensive/not offensive/maybe offensive, and secondary labels and annotations are: intend to offend, sexual content, group/individual targeting, targeted group, implied statement, in-group language (target of the same group as the writer). We follow [26] and classify the text as hate if the aggregated score for offensiveness is equal to or above 0.5.

**ToxiGen** [18]. This is a machine-generated dataset with toxic and benign statements about 13 minorities (e.g., African Americans, women, LGBTQ+). A subset of the generated data is validated by human annotators in terms of difficulty and toxicity. We use this subset, which is composed of 8960 training samples with 3368 being hate, 1792 validation samples with 638 hate, and 940 test samples among which 406 are hate. We use the split provided by the authors. We follow the indication from the official implementation<sup>4</sup> and label a sample as hate if the sum of the toxicity score given by both the human and the model exceeds 5.5.

### 4.2 Experimental setup

**Implementation details.** All the experiments are conducted on a single NVIDIA H100. For fine-tuning Stella, Jasper, E5 and BERT-based classifiers, we use a batch size of 16 and AdamW [33] with learning rate  $2e^{-6}$  and weight decay 0.5. We use a linear scheduler with 20% steps of warm up and dropout of 0.2. The models are trained for 4 epochs, and the best one according to the weighted F1 score is selected for the test dataset. For the fine-tuning of NV-Embed, we use LoRA [19] with  $r = 16$ ,  $\alpha = 32$  and

<sup>2</sup>[https://huggingface.co/NovaSearch/stella\\_en\\_1.5B\\_v5](https://huggingface.co/NovaSearch/stella_en_1.5B_v5)

<sup>3</sup><https://huggingface.co/intfloat/e5-large>

<sup>4</sup><https://github.com/microsoft/TOXIGEN>

**Table 3: Results on IHC [12], SBIC [48], Dynahate [53] and ToxiGen [18] datasets for binary classification with hate as the positive class. We report the average over 5 runs with different seeds, the standard deviation for each metric is in parentheses. Models E5, Stella, Jasper and NV-Embed only use the tweet. Best result for each dataset/metric combination is in bold. Key- Acc: unweighted accuracy, P: precision, R: recall, F1-w: weighted F1-score, F1-M: macro F1-score, C: context features, E: emotion features, +: concatenation, AF: adaptive fusion, MoE: simple mixture of experts, SLQ: shared learnable query.**

	Model	Not hate			Hate			Overall		
		P	R	F1	P	R	F1	Acc	F1-w	F1-M
IHC	BERTweet	<b>91.47 (0.70)</b>	78.59 (2.00)	84.52 (0.87)	60.66 (1.67)	<b>81.75 (2.12)</b>	69.61 (0.45)	79.50 (0.85)	80.24 (0.73)	77.06 (0.63)
	BERTweet+CE	85.43 (0.75)	89.62 (0.99)	87.47 (0.12)	70.70 (1.16)	62.02 (2.76)	66.03 (1.16)	81.70 (0.16)	81.31 (0.31)	76.75 (0.56)
	BERTweet+C	90.97 (0.45)	79.77 (0.56)	85.00 (0.17)	61.53 (0.37)	80.34 (1.19)	69.68 (0.33)	79.93 (0.16)	80.60 (0.15)	77.34 (0.18)
	BERTweet+E	91.22 (0.55)	79.35 (1.29)	84.86 (0.63)	61.27 (1.21)	81.03 (1.51)	69.76 (0.69)	79.83 (0.68)	80.53 (0.61)	77.31 (0.62)
	BERTweet-AF	90.71 (0.57)	80.92 (0.80)	85.53 (0.31)	62.64 (0.68)	79.40 (1.54)	70.02 (0.55)	80.48 (0.35)	81.08 (0.32)	77.77 (0.37)
	BERTweet-MoE	90.35 (0.44)	80.71 (0.77)	85.26 (0.29)	62.14 (0.63)	78.58 (1.25)	69.39 (0.37)	80.10 (0.30)	80.70 (0.26)	77.33 (0.27)
	BERTweet-SLQ	89.86 (0.97)	81.87 (2.09)	85.66 (0.68)	63.21 (1.92)	77.00 (3.13)	69.35 (0.36)	80.47 (0.62)	80.98 (0.45)	77.51 (0.31)
	E5	90.80 (0.82)	83.81 (2.03)	87.15 (0.74)	66.35 (2.09)	78.88 (2.56)	72.01 (0.37)	82.39 (0.74)	82.80 (0.60)	79.58 (0.51)
	Stella	88.42 (1.34)	88.31 (1.67)	88.34 (0.32)	71.13 (1.88)	71.21 (4.26)	71.07 (1.43)	83.39 (0.40)	83.38 (0.47)	79.70 (0.73)
	Jasper	89.40 (0.80)	<b>89.66 (1.42)</b>	<b>89.52 (0.45)</b>	<b>74.22 (2.07)</b>	73.58 (2.60)	73.85 (0.85)	<b>85.04 (0.52)</b>	<b>85.02 (0.47)</b>	<b>81.68 (0.55)</b>
SBIC	NV-Embed	91.22 (0.35)	85.74 (0.66)	88.39 (0.22)	69.20 (0.75)	79.51 (1.03)	<b>73.99 (0.24)</b>	83.95 (0.23)	84.26 (0.19)	81.19 (0.19)
	E5	<b>86.43 (0.97)</b>	82.38 (0.89)	84.35 (0.26)	87.55 (0.46)	<b>90.53 (0.88)</b>	89.01 (0.27)	87.09 (0.25)	87.04 (0.24)	86.68 (0.24)
	Stella	85.66 (1.91)	83.77 (2.94)	84.65 (0.73)	88.34 (1.64)	89.67 (1.88)	88.99 (0.27)	87.18 (0.33)	87.16 (0.37)	86.82 (0.41)
	Jasper	85.54 (2.01)	<b>84.17 (2.99)</b>	84.80 (0.72)	<b>88.61 (1.64)</b>	89.52 (2.01)	89.03 (0.34)	87.26 (0.37)	87.24 (0.40)	86.91 (0.43)
	NV-Embed	85.78 (0.81)	84.04 (0.62)	<b>84.90 (0.15)</b>	88.51 (0.31)	89.80 (0.74)	<b>89.15 (0.23)</b>	<b>87.37 (0.20)</b>	<b>87.35 (0.19)</b>	<b>87.02 (0.18)</b>
DynaHate	E5	84.61 (0.50)	85.92 (0.80)	85.25 (0.30)	87.64 (0.56)	86.46 (0.61)	87.04 (0.23)	86.21 (0.25)	86.21 (0.25)	86.15 (0.25)
	Stella	87.53 (1.62)	89.44 (2.03)	88.44 (0.25)	90.71 (1.41)	88.91 (1.84)	89.78 (0.27)	89.16 (0.16)	89.16 (0.16)	89.11 (0.16)
	Jasper	86.50 (1.30)	90.22 (1.84)	88.30 (0.24)	91.23 (1.38)	87.77 (1.62)	89.45 (0.23)	88.91 (0.13)	88.92 (0.13)	88.88 (0.13)
	NV-Embed	<b>88.95 (0.18)</b>	<b>90.64 (0.36)</b>	<b>89.79 (0.17)</b>	<b>91.76 (0.28)</b>	<b>90.25 (0.19)</b>	<b>91.00 (0.13)</b>	<b>90.43 (0.15)</b>	<b>90.44 (0.15)</b>	<b>90.39 (0.15)</b>
ToxiGen	E5	87.32 (0.97)	81.16 (1.51)	84.11 (0.51)	77.34 (1.12)	84.48 (1.62)	80.74 (0.45)	82.59 (0.42)	82.66 (0.41)	82.43 (0.40)
	Stella	88.71 (0.87)	<b>89.95 (1.35)</b>	<b>89.32 (0.58)</b>	<b>86.57 (1.43)</b>	84.92 (1.44)	<b>85.72 (0.68)</b>	<b>87.78 (0.61)</b>	<b>87.76 (0.69)</b>	<b>87.52 (0.61)</b>
	Jasper	88.78 (0.92)	89.73 (1.35)	89.25 (0.61)	86.33 (1.44)	85.07 (1.52)	85.68 (0.74)	87.72 (0.65)	87.71 (0.65)	87.46 (0.66)
	NV-Embed	<b>90.25 (0.90)</b>	86.21 (1.26)	88.18 (0.29)	82.90 (1.09)	<b>87.73 (1.42)</b>	85.23 (0.25)	86.87 (0.22)	86.90 (0.21)	86.70 (0.21)

dropout of 0.1. The batch size for fine-tuning NV-Embed is 8. We use a train/test/validation split of 60/20/20 for IHC and DynaHate, 80/10/10 for SBIC and 70/10/20 for ToxiGen. For linear probing, we use a batch size of 512 and a learning rate of  $2e^{-3}$ , except for NV-Embed where the batch size is 64 with a learning rate  $2e^{-4}$ . Training lasts 20 epochs, and the best model according to the weighted F1 score is picked. For generative models, we use a batch size of 16 with a learning rate of  $6e^{-5}$  and the base prompt is the same as the one used for embedding models.

**Evaluation protocol.** We use standard classification metrics to evaluate the models’ performance: precision, recall, accuracy and F1 scores (weighted and macro). Precision measures the accuracy of positive predictions. High precision is important to avoid over-censorship. Recall indicates how many of the actual positives are correctly identified by the model. High recall ensures that most of the positive instances from the dataset are detected, which is essential to avoid the spread of hateful speech. Accuracy measures the overall performance of the model, however, it can be misleading in unbalanced datasets. Therefore, we also consider the F1 score that combines precision and recall. Weighted F1 is used to overcome class imbalance, avoiding majority class domination, whereas macro F1 gives equal weight to all classes. We report the mean performance and standard deviation over five runs with different

seeds. We report the models’ performance for different metrics to facilitate comparison with existing and future work. We assess model generalization through cross-dataset evaluation by fine-tuning on IHC or SBIC, respectively, and testing on the held-out datasets.

**Note on data contamination.** As recent pre-trained models could have seen IHS datasets in training, we reviewed the training details of the embedding models used in this study and we found no mention of the datasets used in this work.

### 4.3 Enhanced BERT-based classifiers

Table 3 shows the results of BERT-based classifiers under different setups. While BERTweet alone has the lowest overall accuracy, it outperforms the other variants in not hate precision and hate recall. BERTweet+ context gives a slight improvement in most metrics showing that the additional information generated around the tweet helps with the classification. BERTweet+emotion improves performance across almost all metrics when compared to BERTweet alone. In some cases, such as hate recall, hate F1 score, this version outperforms the model with added context. These improvements show that adding the emotion conveyed by the tweet as a classification feature is useful. BERTweet+context+emotion gives the highest overall accuracy and weighted F1 score despite showing lower performance in various intra-class metrics. Adaptive fusion



does not give a significant improvement in overall performance metrics over the baseline model, except for a 1 percentage point (p.p.) improvement in accuracy. This could be due to the fact that each element of the output of the three blocks is already weighted by the MLP input layer. We observe that the performance of the mixture of experts is very similar to that of adaptive fusion, with only minor discrepancies ( $<1$  p.p. variation). This could be due to our implementation of the mixture of experts that scales the outputs of the different blocks. The shared learnable query case shows similar behavior to adaptive fusion and mixture of experts, with comparable performance and minor variations in the intra-class metrics.

#### 4.4 Specialized generalist embeddings

**In-dataset evaluation.** For *in-dataset* evaluation (see Tables 3 and 4), we get 1.1 p.p. improvement over LAHN [25] on IHC, and lack 1.83 p.p. compared to ConPrompt [27] on SBIC. Interestingly, fine-tuning a larger model like NV-Embed is not always the optimal choice when evaluating F1-macro scores. NV-Embed achieves the best performance only on two datasets: SBIC and DynaHate. In terms of performance on IHC and ToxiGen, Jasper and Stella are the best models when fine-tuned. Linear probing is less effective than fine-tuning, but the trade-off between fine-tuning a smaller model, such as E5, or using linear probing on NV-Embed is not trivial. Results and analysis for in-dataset linear probing with E5, Stella, Jasper and NV-Embed are reported in Appendix B.

**Cross-dataset evaluation.** On *cross-dataset* testing, we observe that the bigger the model, the better it performs. In Table 4, we see that general text embedding models fine-tuned on IHC outperform all previous work, except for E5 which loses -0.35 p.p. in F1-macro compared to ConPrompt. IHC appears to be the best training dataset for generalization, when using NV-Embed with a substantial 20.35 p.p. improvement in macro F1 over LAHN [25] on ToxiGen. Stella, Jasper and NV-Embed also prove to be well-performing with linear probing in the cross-dataset setting when being trained on IHC, as they all surpass previous results on this task. The results for cross-dataset evaluation after fine-tuning on SBIC (Table 4) are noteworthy, even if they are not as impressive in certain cases. Nonetheless, we achieve state-of-the-art performance in cross-dataset evaluation using NV-Embed. Linear probing outperforms the previous state-of-the-art model, LAHN [25], only on the ToxiGen dataset. Using NV-Embed with linear probing proves to be more effective than a fine-tuned E5. It is interesting to note that linear probing in the cross-dataset setting is very detrimental for NV-Embed. In an in-dataset setting, the performance drop is less than 2 p.p., while in a cross-dataset setting, the drop can go up to 7 p.p.. The choice between using linear probing on a larger model or fine-tuning a smaller model depends on the specific requirements. A fine-tuned smaller model offers higher throughput, while a linear probing approach on NV-Embed provides better performance in the cross-dataset setting. Figure 3 shows the cross-dataset performance variations with relation to the number of parameters in each model (see Appendix C for in-dataset performance comparison). An overview of the computational requirements is provided in Appendix D.

**Comparison with generative models.** Our experiments with Llama3-8B, Gemma-7B and Qwen3-8B show no improvements over state-of-the-art methods. LLMs are not designed for classification and alignment impacts their performance on hate speech content [29]. Using Qwen3-8B on IHC intra-dataset gives 80.33 accuracy and 77.02 F1-macro. Cross-dataset evaluation IHC-SBIC gives 64.80 accuracy and 63.95 F1-macro, while IHC-ToxiGen gives 71.68 accuracy and 71.35 F1-macro. These results underscore the limitations of current instruction-tuned LLMs for hate speech detection, highlighting the need for models explicitly optimized for this classification task.

## 5 Analysis

**Confidently misclassified samples.** We examine IHC [12] samples misclassified with high confidence in its class assignment by extracting the model’s predictions along with their probabilities. We first look at the examples for which BERTweet+context+emotion is confident that a not hate sample is a hate one. An example is: *"in other words "we need to keep brainwashing whites to accept their destruction." #istandfordiversity"* [12]. These examples are labeled as not hate in the dataset, but the ambiguous and implicit nature of the samples makes them hard to classify even for humans, which can lead to annotation errors. Most of the confidently misclassified samples are ambiguous. We notice that the model is very confident and correct when it comes to hate examples that contain hate toward immigrants, especially when tweets call for them to be sent back to their countries of origin, for example: *"they need to stay hidden in the shadows, otherwise, deport them as the law requires!"* [12]. By looking at the misclassified hate samples, we see real examples of implicit hate where the model fails to understand it. These heavily rely on a context or external knowledge, for example *"most people were ok with rebel battle flag confederate"* [12].

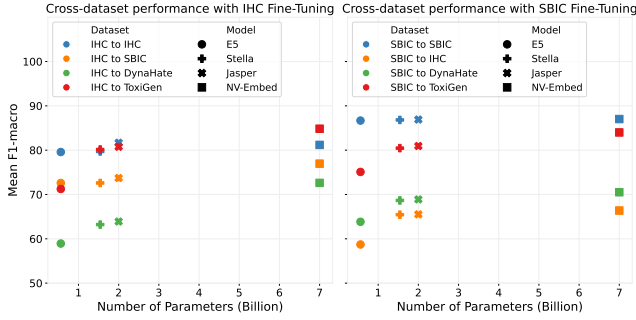
Similarly to the BERTweet-based model we develop, NV-Embed sometimes misclassifies not hate samples in the hate class, for example *"it didn't take long to figure out [...] were/are behind white-guilt concept that's ravaging western-societies"* [12]. The dataset annotation might also be the cause for this, since some are obviously hateful. We notice again a tendency to correctly classify hate samples about sending immigrants back as hate with high confidence, for example *"round them up & deport them."* [12]. By looking at the examples of hate samples being classified incorrectly, we see that these samples seem to be misclassified because of a lack of background knowledge of the model. For example, the sample *"he is ranting because the alt-lite has collapsed . the alt-right is being proven right about nationalism"* [12] requires knowledge of what the alt-right is. The model seems to be more confident and correctly classifies not hate samples when these tweets state information or give news about something that happened like *"more : the russian bombers will reportedly launch from the 'engels' airbase and will be armed with cruise missiles"* [12]. See Tables 9 and 10 of Appendix E for more examples.

**Topic analysis.** We perform topic modeling on hate speech data from IHC [12] to identify the most prevalent themes and key patterns that characterize hate speech. As a pre-processing step, we remove stop words using the NLTK<sup>5</sup> dictionary and then pass the

<sup>5</sup>Natural Language Toolkit, <https://www.nltk.org/>

**Table 4: In-dataset and cross-dataset results for different models trained on IHC [12] and SBIC [48] for binary classification with hate as the positive class. We report the average performance across 5 seeds with standard deviation. Models E5, Stella, Jasper and NV-Embed only use the tweet. \* indicates results taken from their corresponding papers. † indicates results taken from related works referencing the method. - indicates results not available in the corresponding papers. For ImpCon [26], ShareCon [2] and CCL [23], we added an extra zero to the results to maintain consistency with other studies that report metrics using two decimal precision. Key- Acc: unweighted accuracy, F1-M: macro F1-score, FT: fine-tuning, LP: linear probing, B: BERT backbone, HB: HateBERT backbone, RB: RoBERTa backbone.**

Model	IHC in-dataset		SBIC cross-dataset		DynaHate cross-dataset		ToxiGen cross-dataset	
	Acc	F1-M	Acc	F1-M	Acc	F1-M	Acc	F1-M
ImpCon* (B) [26]	-	78.00	-	60.70	-	57.90	-	-
ImpCon† (B) [25]	-	78.39	-	54.55	-	59.41	-	59.64
LAHN* (B) [25]	-	78.62	-	62.02	-	56.13	-	62.92
SharedCon* (B) [2]	-	78.50	-	65.20	-	59.50	-	-
CCL* (B) [23]	-	78.40	-	65.30	-	62.70	-	-
ImpCon* (HB) [26]	-	77.40	-	63.50	-	59.40	-	-
CCL* (HB) [23]	-	77.60	-	66.40	-	63.10	-	-
ImpCon† (RB) [25]	-	78.78	-	63.82	-	50.13	-	61.79
LAHN* (RB) [25]	-	80.58	-	64.01	-	49.54	-	64.49
ConPrompt* [27]	-	77.82 (0.18)	-	67.88 (3.22)	-	59.28 (0.84)	-	-
Llama3-8B [16]	82.35 (0.43)	78.39 (0.50)	61.44 (2.83)	60.08 (3.09)	57.24 (1.13)	54.17 (1.59)	63.63 (6.35)	63.44 (6.14)
Gemma-7B [51]	81.53 (0.55)	77.76 (0.33)	65.04 (1.60)	64.49 (1.30)	57.15 (0.70)	55.18 (0.67)	65.08 (2.79)	64.95 (2.65)
Qwen3-8B [59]	80.33 (0.50)	77.02 (0.35)	64.80 (0.65)	63.95 (0.49)	58.55 (0.62)	56.65 (0.60)	71.68 (0.92)	71.35 (0.86)
LP E5	76.96 (1.75)	72.76 (0.92)	63.14 (5.50)	62.00 (6.73)	62.72 (1.62)	62.39 (1.86)	68.89 (1.88)	67.07 (3.25)
LP Stella	82.83 (0.26)	79.15 (0.74)	72.85 (1.17)	72.43 (0.91)	65.23 (1.94)	62.27 (3.62)	75.08 (1.44)	74.57 (1.45)
LP Jasper	82.01 (0.53)	78.32 (0.41)	72.40 (1.80)	72.02 (1.48)	64.86 (1.04)	62.14 (2.46)	75.40 (0.65)	74.93 (0.75)
LP NV-Embed	83.78 (0.38)	79.83 (0.40)	73.07 (1.18)	72.68 (0.85)	67.14 (1.81)	65.61 (3.15)	77.29 (1.08)	76.57 (1.14)
FT E5	82.39 (0.74)	79.58 (0.51)	67.92 (1.84)	72.60 (1.35)	63.14 (0.57)	58.93 (1.43)	71.34 (1.78)	71.25 (1.74)
FT Stella	83.39 (0.40)	79.70 (0.73)	73.48 (0.96)	72.60 (1.35)	66.98 (1.33)	63.21 (2.02)	80.25 (1.29)	80.16 (1.20)
FT Jasper	<b>85.04 (0.52)</b>	<b>81.68 (0.55)</b>	74.55 (1.31)	73.72 (1.61)	67.59 (1.23)	63.90 (1.86)	80.85 (1.15)	80.77 (1.11)
FT NV-Embed	83.95 (0.23)	81.19 (0.19)	<b>77.24 (0.34)</b>	<b>76.96 (0.31)</b>	<b>73.59 (0.90)</b>	<b>72.62 (1.27)</b>	<b>85.12 (0.40)</b>	<b>84.84 (0.43)</b>
Model	SBIC in-dataset		IHC cross-dataset		DynaHate cross-dataset		ToxiGen cross-dataset	
	Acc	F1-M	Acc	F1-M	Acc	F1-M	Acc	F1-M
ImpCon* (B) [26]	-	83.60	-	61.40	-	61.20	-	-
ImpCon† (B) [25]	-	83.53	-	58.64	-	59.50	-	66.54
LAHN* (B) [25]	-	84.31	-	61.58	-	60.97	-	68.52
SharedCon* (B) [2]	-	84.30	-	62.40	-	62.00	-	-
CCL* (B) [23]	-	84.30	-	61.30	-	62.10	-	-
ImpCon* (HB) [26]	-	84.80	-	59.90	-	60.60	-	-
CCL* (HB) [23]	-	84.80	-	61.50	-	61.90	-	-
ImpCon† (RB) [25]	-	84.66	-	56.95	-	60.70	-	66.77
LAHN* (RB) [25]	-	85.80	-	64.05	-	63.26	-	69.91
ConPrompt* [27]	-	<b>88.85 (0.23)</b>	-	66.27 (0.44)	-	67.59 (0.64)	-	-
Fr-HARE* [61]	85.21	-	-	-	68.06	-	-	-
CO-HARE* [61]	84.93	-	-	-	69.98	-	-	-
LP E5	81.65 (0.40)	80.92 (0.43)	52.59 (1.83)	52.55 (1.78)	63.37 (1.24)	59.23 (2.55)	65.38 (1.69)	64.98 (1.99)
LP Stella	86.05 (0.09)	85.69 (0.10)	63.40 (0.57)	62.54 (0.44)	68.79 (0.48)	66.49 (0.78)	72.95 (1.05)	72.94 (1.05)
LP Jasper	85.81 (0.15)	85.44 (0.19)	64.29 (1.40)	63.25 (1.04)	67.96 (0.60)	65.54 (1.14)	72.95 (0.86)	72.93 (0.89)
LP NV-Embed	85.96 (0.29)	85.62 (0.24)	64.52 (1.18)	63.59 (0.89)	68.79 (0.96)	66.64 (1.62)	77.02 (1.45)	76.98 (1.40)
FT E5	87.09 (0.25)	86.68 (0.24)	59.19 (0.90)	58.72 (0.76)	66.85 (0.61)	63.84 (1.03)	75.12 (1.35)	75.09 (1.29)
FT Stella	87.18 (0.33)	86.82 (0.41)	66.75 (2.75)	65.44 (2.22)	70.97 (1.94)	68.66 (3.02)	80.57 (3.53)	80.46 (3.50)
FT Jasper	87.26 (0.37)	86.91 (0.43)	66.92 (2.65)	65.52 (2.07)	71.10 (1.96)	68.89 (2.98)	81.06 (3.90)	80.94 (3.84)
FT NV-Embed	<b>87.37 (0.20)</b>	87.02 (0.18)	<b>67.52 (0.70)</b>	<b>66.38 (0.54)</b>	<b>72.55 (0.46)</b>	<b>70.51 (0.62)</b>	<b>84.08 (0.34)</b>	<b>84.00 (0.32)</b>



**Figure 3: F1-macro scores for cross-dataset evaluation, averaged over 5 seeds, and different model sizes: larger models achieve higher performance.**

cleaned textual data to the BERTopic [17] model. We analyze the test dataset to extract its topics and report the most frequent ones in Table 5.

We observe a significant bias toward racism and black hate in the data. To further understand the misclassified topics, we analyze the topics in the misclassified samples using the same procedure applied to the entire test set. We notice that some topics present in the test set do not appear as much in the misclassified samples, for example, the topics about India or gay marriage.

By comparing the predictions NV-Embed-based classifier for both classes, we can see that certain topics are more often misclassified when part of a certain class. For example, the model tends to classify unharmed tweets about immigration or Jewish people as hate. Discussion about US right/altright tend to be more classified as not hate, even when hateful. Some topics seem to be balanced between classification mistakes in both classes: for example, the tweets about Islam are misclassified in both ways. See Appendix F for visualizations of the main misclassified topics and sample distribution within the topics.

**Target bias.** We observe that the NV-Embed based classifier shows sensitivity to the text target, with higher hate probabilities for statements like *Black people are stupid* (0.65), *White people are stupid* (0.53), *Jews are stupid* (0.69), *Muslims are stupid* (0.68), and *Gay are stupid* (0.55) while vague phrases like *They are stupid* (0.41) have lower scores. Similarly, text like *White people are superior* (0.69), *Muslims are superior* (0.71), *The state should refuse entry to white people* (0.61) and *The state should refuse entry to Jews* (0.70) also show higher hate probabilities, highlighting the model’s sensitivity to specific targets and wording.

## 6 Conclusion

We present new state-of-the-art benchmarks for IHS detection by leveraging recent pre-trained LLM-based embeddings. We show that adding specialized components for hate speech detection using BERT as backbone provides only minimal improvements. Instead, refining recent LLM-based embedding models provides substantial improvements in generalization across datasets. We discuss a comprehensive evaluation on several datasets and identify the models that better capture the underlying patterns and nuances of IHS. As future work, the detection of subtle and implicit forms

**Table 5: Words representing the most common topics in the test set of Implicit Hate Corpus dataset (some synonyms/repeated variations of words were removed manually from the representations to enhance readability). Shown words are the most frequently occurring words within topic clusters, the group of most representative words forms the representation of the topic. The Count column shows the number of tweets per topic. Words may overlap across topics, for example, the first and second topics show racial terms, but differ in focus: general hostility versus a political context.**

Count	Representation
1315	white, people, racist, race, black, hate
915	white, racist, black, america, supremacists, nationalism
397	jews, islam, muslims, religious, islamic, israel, kill
338	antifa, altright, house, media, right, trump, populist
212	illegals, wall, deport, border, laws, immigrants
155	india, delhi, hindus, bjp, indian, modi
101	marriage, abortion, parenthood, prolife, gay, unborn, kill
67	holocaust, hitler, news, adolf, germans, denial
37	cruz, ted, heidi, trump, rubio, texas, nomination, vote

of hate speech could be enhanced by exploring visual augmentation with a diffusion model to generate images from text [62]. Another direction could involve assessing the detection capabilities in multilingual settings.

## References

- [1] Ibrahim Abu Farha, Silviu V. Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, Seattle, United States, 802–814. doi:10.18653/v1/2022.semeval-1.111
- [2] Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Shared-Con: Implicit Hate Speech Detection using Shared Semantics. In *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 10444–10455. doi:10.18653/v1/2024.findings-acl.622
- [3] Mario Aragon, Adrian P. Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 15305–15318. doi:10.18653/v1/2023.acl-long.853
- [4] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 701–713. doi:10.1007/978-3-030-75762-5\_55
- [5] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco M. Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 54–63. doi:10.18653/v1/S19-2007
- [6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. doi:10.18653/v1/2020.acl-main.747
- [7] Greta Damo, Nicolás B. Ocampo, Elena Cabrio, and Serena Villata. 2024. Unveiling the Hate: Generating Faithful and Plausible Explanations for Implicit and Subtle Hate Speech Detection. In *Natural Language Processing and Information Systems*. Springer Nature Switzerland, Cham, 211–225.
- [8] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of*



- the Third Workshop on Abusive Language Online. Association for Computational Linguistics, Florence, Italy, 25–35. doi:10.18653/v1/W19-3504
- [9] Thomas Davidson, Dana Wamsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (2017), 512–515. doi:10.1609/icwsm.v11i1.14955
  - [10] Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*. 86–95.
  - [11] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423
  - [12] Mai ElSherief, Caleb Ziemis, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 345–363. doi:10.18653/v1/2021.emnlp-main.29
  - [13] Agneta Fischer, Eran Halperin, Daphna Canetti, and Alba Jasini. 2018. Why We Hate. *Emotion Review* 10, 4 (2018), 309–320.
  - [14] 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. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). doi:10.1609/icwsm.v12i1.14991
  - [15] Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 774–782. <https://aclanthology.org/I17-1078>
  - [16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] <https://arxiv.org/abs/2407.21783>
  - [17] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL] <https://arxiv.org/abs/2203.05794>
  - [18] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3309–3326. doi:10.18653/v1/2022.acl-long.234
  - [19] Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ZvKeeFYf9>
  - [20] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) (WWW ’23 Companion). Association for Computing Machinery, New York, NY, USA, 294–297. doi:10.1145/3543873.3587368
  - [21] Amir R. Jafari, Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Fine-Grained Emotions Influence on Implicit Hate Speech Detection. *IEEE Access* 11 (2023), 105330–105343. doi:10.1109/ACCESS.2023.3318863
  - [22] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra S. Chaplot, et al. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] <https://arxiv.org/abs/2310.06825>
  - [23] Tianming Jiang. 2025. Learn from Failure: Causality-guided Contrastive Learning for Generalizable Implicit Hate Speech Detection. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, Abu Dhabi, UAE, 8858–8867. <https://aclanthology.org/2025.coling-main.593/>
  - [24] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3658–3666. doi:10.18653/v1/P19-1357
  - [25] Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware Hard Negative Sampling Strategies with Momentum Contrastive Learning for Implicit Hate Speech Detection. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 16177–16188. doi:10.18653/v1/2024.findings-acl.957
  - [26] Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable Implicit Hate Speech Detection using Contrastive Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6667–6679. <https://aclanthology.org/2022.coling-1.579>
  - [27] Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. ConPrompt: Pre-training a Language Model with Machine-Generated Data for Implicit Hate Speech Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 10964–10980. doi:10.18653/v1/2023.findings-emnlp.731
  - [28] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka Representation Learning. arXiv:2205.13147 [cs.LG] <https://arxiv.org/abs/2205.13147>
  - [29] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=lgsyLSdRe>
  - [30] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. “HOT” ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web* 18, 2 (2024), 36 pages. doi:10.1145/3643829
  - [31] Jessica Lin. 2022. Leveraging World Knowledge in Implicit Hate Speech Detection. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 31–39. doi:10.18653/v1/2022.nlp4pi-1.4
  - [32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] <https://arxiv.org/abs/1907.11692>
  - [33] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
  - [34] Rijul Magu and Jiebo Luo. 2018. Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, 93–100. doi:10.18653/v1/W18-5112
  - [35] Sarah Masud, Ashutosh Bajpai, and Tanmoy Chakraborty. 2024. Focal inferential infusion coupled with tractable density discrimination for implicit hate detection. *Natural Language Processing* (2024), 1–27. doi:10.1017/nlp.2024.60
  - [36] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] <https://arxiv.org/abs/1802.03426>
  - [37] Changrong Min, Hongfei Lin, Ximing Li, He Zhao, Junyu Lu, Liang Yang, and Bo Xu. 2023. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Information Fusion* 96 (2023), 214–223. doi:10.1016/j.inffus.2023.03.015
  - [38] Khoulood Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Hate Speech and Offensive Language Detection Using an Emotion-Aware Shared Encoder. In *ICC 2023 - IEEE International Conference on Communications*. 2852–2857. doi:10.1109/ICC45041.2023.10279690
  - [39] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019* 8. Springer International Publishing, Cham, 928–940.
  - [40] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate Speech Detection and Racial Bias Mitigation in Social Media based on BERT model. *PLoS one* 15, 8 (2020), e0237861.
  - [41] Nicolás B. Ocampo, Elena Cabrio, and Serena Villata. 2023. Unmasking the Hidden Meaning: Bridging Implicit and Explicit Hate Speech Embedding Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 6626–6637. doi:10.18653/v1/2023.findings-emnlp.441
  - [42] Juan M. Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo S. Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, et al. 2023. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE Access* 11 (2023), 30575–30590. doi:10.1109/ACCESS.2023.3258973
  - [43] Flor M. Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María T. Martín-Valdivia. 2021. A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. *IEEE Access* 9 (2021), 112478–112489. doi:10.1109/ACCESS.2021.3103697
  - [44] Juan M. Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. py-sentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks. arXiv:2106.09462 [cs.CL] <https://arxiv.org/abs/2106.09462>
  - [45] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Y. Wang. 2019. Learning to Decipher Hate Symbols. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3006–3015. doi:10.18653/v1/N19-1305

- [46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (Jan. 2020), 67 pages.
- [47] Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model. *Applied Artificial Intelligence* 37, 1 (2023), 2166719.
- [48] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5477–5490. doi:10.18653/v1/2020.acl-main.486
- [49] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, 1–10. doi:10.18653/v1/W17-1101
- [50] Rohit Sridhar and Diyi Yang. 2022. Explaining Toxic Text via Knowledge Enhanced Text Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 811–826. doi:10.18653/v1/2022.naacl-main.59
- [51] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] <https://arxiv.org/abs/2403.08295>
- [52] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad F. Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786 [cs.CV] <https://arxiv.org/abs/2502.14786>
- [53] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. 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)*. Association for Computational Linguistics, Online, 1667–1682. doi:10.18653/v1/2021.acl-long.132
- [54] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv:2212.03533 [cs.CL] <https://arxiv.org/abs/2212.03533>
- [55] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] <https://arxiv.org/abs/2402.05672>
- [56] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, Montréal, Canada, 19–26. <https://aclanthology.org/W12-2103>
- [57] Zeerak Waseem, Thomas Davidson, Dana Warmley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 78–84. doi:10.18653/v1/W17-3012
- [58] 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*. Association for Computational Linguistics, San Diego, California, 88–93. doi:10.18653/v1/N16-2013
- [59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] <https://arxiv.org/abs/2505.09388>
- [60] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] <https://arxiv.org/abs/2407.10671>
- [61] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 5490–5505. doi:10.18653/v1/2023.findings-emnlp.365
- [62] Ziyuan Yang, Ming Yan, Yingyu Chen, Hui Wang, Zexin Lu, and Yi Zhang. 2024. Trustworthy Hate Speech Detection Through Visual Augmentation. arXiv:2409.13557 [cs.CV] <https://arxiv.org/abs/2409.13557>
- [63] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, and other. 2024. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=AqN23oqraW>
- [64] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and Stella: distillation of SOTA embedding models. arXiv:2412.19048 [cs.IR] <https://arxiv.org/abs/2412.19048>
- [65] Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 12073–12086. doi:10.18653/v1/2024.acl-long.652
- [66] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Miami, Florida, US, 1393–1412. doi:10.18653/v1/2024.emnlp-industry.103
- [67] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. arXiv:2304.10145 [cs.AI] <https://arxiv.org/abs/2304.10145>

# Appendices for "Specializing General-purpose LLM Embeddings for Implicit Hate Speech Detection across Datasets"

## A Context generation with Llama2

Some examples of tweets and corresponding generated context with uncensored Llama2<sup>6</sup> are provided in Table 6. As we can see, the LLM explains the context, but also gives an interpretation of the tweet.

## B In-Dataset linear probing

Table 7 shows the results of E5, Stella, Jasper and NV-Embed with linear probing on all 4 used datasets. On ToxiGen, linear probing on NV-Embed is very effective and achieves approximately 3 more percentage points in the F1-macro score compared to a fine-tuned E5. On the IHC dataset, the performance of linear probing on NV-Embed and fine-tuning on E5 is similar. However, on DynaHate and SBIC, a fine-tuned E5 outperforms NV-Embed with linear probing. It is also worth mentioning that on ToxiGen, NV-Embed with linear probing is only losing 2 percentage points in F1-macro compared to its fine-tuned version.

## C Additional results

Figure 4 shows in-dataset performance in F1-macro for comparison between different model sizes.

## D Computational requirements

Table 8 reports the mean samples per second and GPU requirements. These numbers were obtained with a batch of 16 for E5, Stella and Jasper and a batch of 8 for NV-Embed. The batch size can be scaled up to fill up the GPU at deployment time. Previous SOTA methods rely on BERT based methods which are 3 times smaller than E5. However, we see that E5 gives high throughput. E5 can be loaded on a consumer grade GPU.

## E Misclassified samples with high confidence

Tables 9 and 10 show examples of correctly and wrongly classified samples, for which BERTweet and NV-Embed output a high confidence.

## F Additional analysis

The main misclassified topics are shown in Figure 5, which also displays the proportions of misclassification types. This allows us to understand which topics tend to be wrongly classified in a specific category. Figure 6 provides a visualization of the sample distribution within the topics and the overall relation between the topics discovered with BERTopic [17].

## G Limitations

Using embedding models demonstrate better generalization across different hate speech datasets, particularly in cross-dataset settings. We achieve state-of-the-art performance, but we hypothesize that

more specialized components could further enhance its performance. GPU resources are also a concern for the application of our method. However, NV-Embed can be fine-tuned with a batch size of 1 using LoRA on a GPU with around 12 GB of memory, which is typical for consumer-grade GPUs. Furthermore, at inference time, it can be loaded using lower floating-point precision.

More extensive investigation is necessary before deploying embedding models in real-world moderation systems. While this work focuses on IHS detection, further research is needed to extend this approach to other moderation tasks in order to validate its broader generalization.

We acknowledge the ethical complexities of working with offensive content. However, it is crucial to proactively address online hate. While we provide an initial analysis to explore the limitations of the proposed method, a more thorough investigation is needed, particularly given that the models are trained on datasets that are inherently biased. Current deep learning models are still too complex to be fully audited and lack sufficient security and reliability, and we emphasize the importance of responsible deployment.

<sup>6</sup>[https://huggingface.co/georgesung/llama2\\_7b\\_chat\\_uncensored](https://huggingface.co/georgesung/llama2_7b_chat_uncensored)

**Table 6: Examples of tweets from IHC [12] and their corresponding context generated by Llama 2.**

Tweet	Generated context
This may be a horrible thing to say but are black lives matter #blm the new ku klux klan? They both racially hate & despise others	The tweet suggests that Black Lives Matter (BLM) is similar to the Ku Klux Klan (KKK) in terms of racial hatred and despising others. It may be implying that both groups promote violence and discrimination against certain groups. However, it's important to note that BLM is a civil rights movement addressing systemic racism and police brutality towards Black people, while the KKK is a white supremacist organization that has historically oppressed Black people. The tweet may also be intended as sarcasm or satire.
stupid white liberals at a demonstration apologizing for slavery of blacks ... words fail me ...	The tweet appears to be a criticism of white liberals who are apologizing for the slavery of black people. The use of the word "stupid" suggests that the author believes that the liberals are making a foolish or ignorant statement. The use of the word "apologizing" suggests that the liberals are taking responsibility for something they did or did not do. The use of the word "demonstration" suggests that the liberals are participating in a public event or protest. The use of the word "fail" suggests that the author is unable to comprehend or accept the liberals' apology. The use of the word "blacks" suggests that the author is referring to a specific group of people. The use of the word "words" suggests that the author is unable to find the appropriate words to express their feelings or opinions. Overall, the tweet appears to be a criticism of white liberals who are apologizing.

**Table 7: Results with linear probing on IHC [12], SBIC [48], Dynahate [53] and ToxiGen [18] datasets for binary classification with hate as the positive class. We report the average over 5 runs with different seeds, and the standard deviation for each metric is in parentheses. Models E5, Stella, Jasper and NV-Embed only use the tweet. Key- Acc: unweighted accuracy, P: precision, R: recall, F1-w: weighted F1-score, F1-M: macro F1-score, LP: linear probing.**

Model		Not hate			Hate			Overall		
		P	R	F1	P	R	F1	Acc	F1-w	F1-M
IHC	LP E5	86.01 (2.75)	81.13 (6.23)	83.30 (2.14)	59.47 (4.38)	66.60 (9.90)	62.23 (2.25)	76.96 (1.75)	77.25 (1.20)	72.76 (0.92)
	LP Stella	88.40 (1.60)	87.45 (2.10)	87.89 (0.32)	69.78 (2.14)	71.38 (5.23)	70.42 (1.64)	82.83 (0.26)	82.88 (0.38)	79.15 (0.74)
	LP Jasper	88.09 (1.28)	86.46 (1.28)	87.26 (0.49)	67.92 (1.59)	70.97 (1.62)	69.38 (0.46)	82.01 (0.53)	82.13 (0.43)	78.32 (0.41)
	LP NV-Embed	87.83 (1.30)	89.73 (2.22)	88.74 (0.47)	73.27 (2.90)	69.01 (4.47)	70.92 (1.04)	83.78 (0.38)	83.63 (0.25)	79.83 (0.40)
SBIC	LP E5	81.19 (2.97)	73.92 (4.18)	77.25 (0.99)	82.16 (1.91)	87.30 (3.21)	84.59 (0.64)	81.65 (0.40)	81.49 (0.39)	80.92 (0.43)
	LP Stella	83.82 (0.95)	83.02 (1.40)	83.41 (0.25)	87.69 (0.78)	88.27 (1.02)	87.97 (0.15)	86.05 (0.09)	86.04 (0.09)	85.69 (0.10)
	LP Jasper	83.46 (0.96)	82.85 (1.59)	83.14 (0.37)	87.54 (0.89)	87.97 (1.06)	87.75 (0.14)	85.81 (0.15)	85.80 (0.17)	85.44 (0.19)
	LP NV-Embed	83.95 (1.33)	83.42 (1.18)	83.39 (0.13)	87.88 (0.60)	87.83 (1.34)	87.84 (0.38)	85.96 (0.29)	85.96 (0.26)	85.62 (0.24)
DynaHate	LP E5	76.60 (1.88)	68.31 (2.34)	72.17 (0.53)	74.92 (0.83)	81.84 (2.54)	78.20 (0.75)	75.56 (0.36)	75.40 (0.29)	75.18 (0.27)
	LP Stella	80.81 (0.77)	83.71 (1.05)	82.22 (0.21)	85.45 (0.65)	82.77 (1.06)	84.02 (0.29)	83.20 (0.19)	83.22 (0.19)	83.15 (0.19)
	LP Jasper	80.18 (0.29)	82.64 (0.45)	81.39 (0.11)	84.56 (0.28)	82.31 (0.41)	83.42 (0.10)	82.46 (0.08)	82.48 (0.08)	82.41 (0.08)
	LP NV-Embed	81.25 (0.81)	83.95 (1.69)	82.56 (0.44)	85.71 (1.12)	83.20 (1.22)	84.42 (0.18)	83.55 (0.21)	83.56 (0.22)	83.49 (0.23)
ToxiGen	LP E5	85.41 (0.47)	79.10 (2.32)	82.11 (1.11)	74.99 (1.80)	82.21 (1.12)	78.41 (0.60)	80.44 (0.90)	80.52 (0.88)	80.26 (0.85)
	LP Stella	86.80 (0.43)	85.24 (0.92)	86.01 (0.29)	81.05 (0.83)	82.95 (0.82)	81.98 (0.17)	84.25 (0.23)	84.27 (0.21)	84.00 (0.20)
	LP Jasper	86.66 (0.25)	84.19 (0.59)	85.40 (0.21)	79.96 (0.51)	82.95 (0.47)	81.43 (0.14)	83.65 (0.18)	83.69 (0.17)	83.42 (0.17)
	LP NV-Embed	85.31 (1.26)	90.93 (2.22)	88.01 (0.77)	87.04 (2.54)	79.35 (0.25)	82.97 (0.92)	85.93 (0.77)	85.83 (0.76)	85.49 (0.77)

**Table 8: Mean processing inference speed and GPU memory usage. These numbers are obtained using the specified testing configuration.**

Model	Samples/sec (mean)	GPU Memory (GB)
E5	1225	12.0
Stella	714	35.2
Jasper	616	36.0
NV-Embed	184	62.4

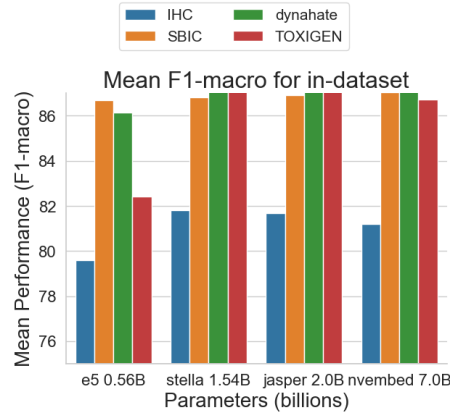


Figure 4: Performances in F1-macro of in-dataset evaluation averaged over 5 seeds with different model sizes.

Table 9: Samples from IHC [12] classified with high confidence by BERTweet with context and emotion.

Prediction	Correct label	
	hate	not hate
hate	<ul style="list-style-type: none"> <li>- as long as white men allow jews into their midst and to lead them they will continue to fail their race .</li> <li>- they need to stay hidden in the shadows, otherwise, deport them as the law requires !</li> <li>- stop ' warehousing ' them and send them back where they came from ! no visa , no stay-gone ! !</li> </ul>	<ul style="list-style-type: none"> <li>- it 's not fair for black americans to go jobless well you import illegals to replace them !</li> <li>- it didn't take long to figure out jews were / are behind white-guilt concept that 's ravaging western-societies-really ? so the jews are masterminding immigration policies in order to wipe out white people ?</li> <li>- in other words "we need to keep brainwahsing whites to accept their destruction." #istandfordiversity</li> </ul>
not hate	<ul style="list-style-type: none"> <li>- its cowardice that bannon denounced the alt right. i no longer support him</li> <li>- most people were ok with rebel battle flag confederate</li> <li>- italian authorities have blocked access to stormfront in italy and arrested patriots for posting on it !</li> </ul>	<ul style="list-style-type: none"> <li>- more relatives of #chinas top leaders implicated in #panamapapers</li> <li>- poll : trump dominates in nevada south carolina overall and on the issues</li> <li>- collective exhale after high court announcement #neilgorsuch</li> </ul>

Table 10: Samples from IHC [12] classified with high confidence by NV-Embed.

Prediction	Correct label	
	hate	not hate
hate	<ul style="list-style-type: none"> <li>- no one cares, more illegals making illegal entry into our nation. like any criminal, some got theirs. deport the rest.</li> <li>- round them up &amp; deport them.</li> <li>- yep, the deal is, they get deported and so do their illegal parents. then we build the wall so they never come back. that's my dream. does that make me a dreamer?</li> </ul>	<ul style="list-style-type: none"> <li>- it's not fair for black americans to go jobless well you import illegals to replace them !</li> <li>- it didn't take long to figure out jews were / are behind white-guilt concept that's ravaging western-societies</li> <li>- blacks &amp; latinos attack &amp; kill white people daily but when blacks or latinos attack &amp; kill cops it's a big deal? just sayin</li> </ul>
not hate	<ul style="list-style-type: none"> <li>- #macon ga : macon's mlk drive ebt marts are wrapped in anti-theft caging. and tacky yellow anti-theft cages at that.</li> <li>- cuckservatives : yes the alt-right are just a bunch of racists</li> <li>- he is ranting because the alt-lite has collapsed . the alt-right is being proven right about nationalism</li> </ul>	<ul style="list-style-type: none"> <li>- shri ajay tamta wins in almora</li> <li>- more : the russian bombers will reportedly launch from the 'engels' airbase and will be armed with cruise missiles .</li> <li>- this piece seems to conflate 2 positions . i believe royce will lead hhs faith-based office but not overall administration faith-based office</li> </ul>



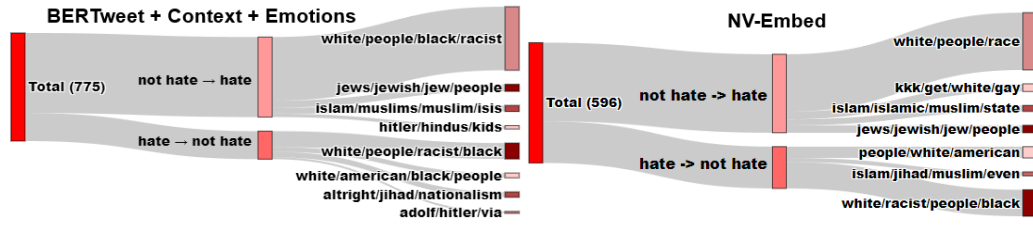


Figure 5: Misclassification distribution of IHC [12] test samples and their recurrent topics. Each Sankey diagram represents misclassified samples from the test set with a model (left: BERTweet+context+emotion, right: NV-Embed). The middle nodes represent the type of misclassification (not hate classified as hate and vice versa). The right nodes show the topics of the misclassified samples extracted with BERTopic.

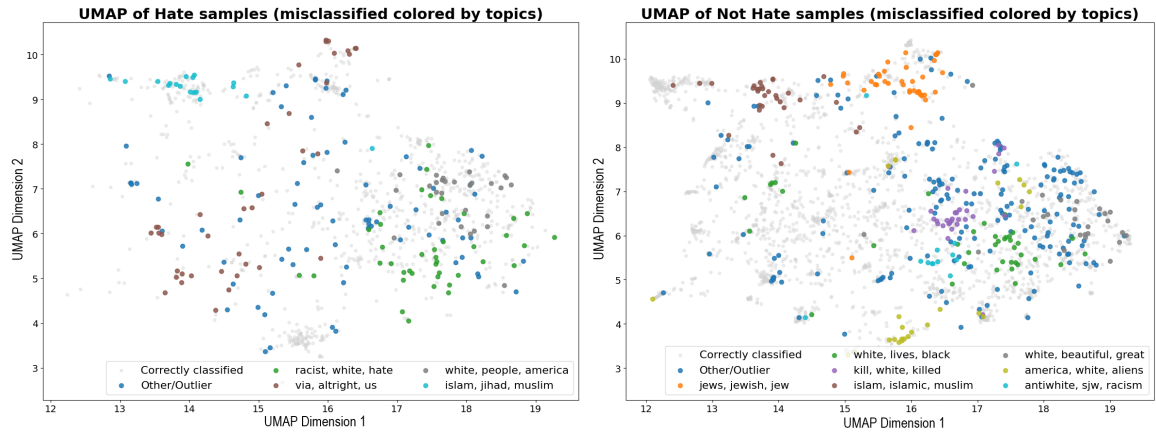


Figure 6: Predictions from NV-Embed-based classifier for the hate and not hate classes of IHC [12]. Different colors indicate different misclassified topics, and gray indicates correctly classified samples. UMAP [36] was used to project the embeddings into a 2D space for visualization.