# Data-Efficient Methods For Improving Hate Speech Detection

**Sumegh Roychowdhury** *
IIT Kharagpur, India
sumeghtech@gmail.com

**Vikram Gupta**
ShareChat, India
vikramgupta@sharechat.co

## Abstract

Scarcity of large-scale datasets, especially for resource-impoverished languages encouraged exploration of data-efficient methods for hate speech detection. In this work, we progress *implicit* and *explicit* hate speech detection using an input-level data augmentation technique, task reformulation using entailment and cross-learning across five languages. Our proposed data augmentation technique `EasyMixup`, improves the F1 performance across languages by **0.5-9%**. We also observe substantial F1 gains of **1-8%** by reformulating hate speech detection as `Entailment-style` problem. We further probe the contextual models and observe that higher layers encode *implicit* hate while lower layers focus on *explicit hate*, highlighting the importance of token-level understanding for *explicit* and context-level for *implicit* hate speech detection. [1]

## 1 Introduction

Deep learning based methods (Badjatiya et al., 2017; Zhang et al., 2018; Kshirsagar et al., 2018) have shown impressive results in detecting hate speech. Transformer based models (Caselli et al., 2021; Tekiroğlu et al., 2020; Aluru et al., 2020; Mozafari et al., 2019; Dutta et al., 2022) have further pushed the state-of-the-art by leveraging large amount of unlabeled data in a self-supervised manner. Various hate speech detection datasets have been contributed in textual (Gibert et al., 2018; Davidson et al., 2017; Founta et al., 2018), audio (Gupta et al., 2022) and visual (Gomez et al., 2020) domains. However, these algorithms are data-hungry and motivate development of algorithms which are data-efficient.

To tackle this, we introduce an input-level data augmentation technique `EasyMixup` and improve hate speech detection in monolingual and multilingual settings. `EasyMixup` is inspired by *mixup* based augmentation techniques which are broadly categorized into input-level mixup (Yun et al., 2019; Kim et al., 2020; Uddin et al., 2021; Walawalkar et al., 2020) and hidden-level mixup (Verma et al., 2019). `EasyMixup` follows the input-level paradigm and leverages a simple observation that the label of a hateful instance is preserved on concatenation with a hateful or non-hateful instance. Similarly, label of a non-hateful instance does not change on concatenation with another non-hateful instance.

We also study the efficacy of reformulating hate speech detection as `Entailment-style` problem. We extend the work by (Wang et al., 2021) and perform detailed experiments under *implicit*, *explicit* and *multilingual* settings. We observe that monolingual entailment performs better than English based entailment. This observation is intuitive because the models are pretrained using pair of sentences from same language and monolingual entailment reflects the same settings.

Majority of the existing textual datasets focus on *explicit* hate speech where *swear, cuss, abusive* words are used to express the hateful intent. In contrast, *implicit* hate speech employs subtle, indirect and contextual ways for expressing hate speech making it extremely harmful and difficult, as shown in (ElSherief et al., 2021). Acknowledging this difference of expression, we explore the relationship between *explicit* and *implicit* hate speech using cross-learning and observe strong correlations. We also perform probing experiments and observe that lower layers focus on *explicit* hate speech while higher layers are responsible for encoding *implicit* hate speech. This alludes to the hypothesis that *implicit* hate speech is more contextual in nature and requires more understanding, while *explicit* hate speech can be detected by leveraging lower-level information.

In summary, our main contributions are:

* Work done during internship at ShareChat

[1] Code and Dataset splits - https://github.com/Sumegh-git/data_efficient_hatedetect



*Findings of the Association for Computational Linguistics: EACL 2023*, pages 125–132
May 2-6, 2023 

- We propose input-level data augmentation technique `EasyMixup` which outperforms previous methods for our task.

- We show performance gains by reformulating hate speech detection as monolingual `Entailment-style` problem.

- We probe contextual models and observe that higher layers encode *implicit* hate speech while lower layers focus on *explicit hate* speech.

- We show that correlations exist between *explicit* and *implicit* hate speech and leverage that for improving hate speech detection.

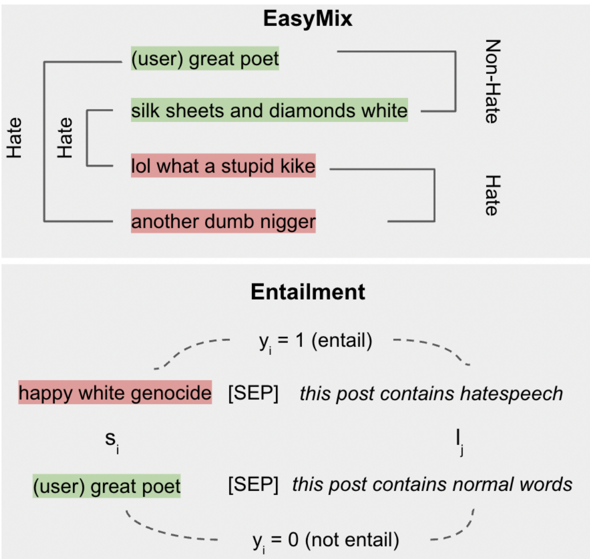


Figure 1: *(top)* Mixing hateful (red) samples with either hate (red) or non-hate (green) samples doesn't change the final label. Similarly mixing two non-hate (green) samples preserve the final label. *(bottom)* Posing hatespeech classification as entailment task. [Best viewed in color]

# 2 Methodology

## 2.1 EasyMixup

`EasyMixup` is an input-level data augmentation technique where we leverage the observation that ground truth label of a hateful sample does not change on concatenation with another hateful or non-hateful samples. Similarly, concatenation of a non-hate sample with another non-hate sample results in a novel non-hate sample as shown in Figure 1(top). More formally, let's say $(s_i, y_i)$ is the sentence and it's corresponding label $y \in \{$*hate*, *non-hate*$\}$ in a minibatch $S$ and $D$ is the entire dataset,

$$S = \{(s_0, y_0), (s_1, y_1), ..., (s_n, y_n) | (s_i, y_i) \in D\}$$

For every sample in the batch, $s_i \in S$, we randomly select $(\overline{s_i}, \overline{y_i}) \in D$ with $\overline{s_i} \neq s_i$ and augmentation probability $p_{aug}$ to create new augmented sample:

$$s_{i_{aug}} = \phi(s_i, \overline{s_i}), y_{i_{aug}} = y_i \vee \overline{y_i}$$

where $\phi$ is defined as :

$$\phi(s_i, \overline{s_i}) = \begin{cases} concat(s_i; \overline{s_i}) & \overline{p} > p_{flip} \\ concat(\overline{s_i}; s_i) & , otherwise \end{cases}$$

where, $p_{flip}$ is the sentence flipping probability and $concat()$ refers to concatenation. Flipping introduces more augmentation and prevents the model from learning positional bias. Finally, we get the updated minibatch $\overline{S}$ by replacing original with augmented samples $(s_{i_{aug}}, y_{i_{aug}})$.

## 2.2 Entailment-style

We reformulate hate speech classification task as an entailment-style task (Wang et al., 2021). The (*input*, *target*) for the contextual model is: $(s_i[sep]l_j, y_i)$, where, $s_i$ is the original sentence, $l_j$ is the label-prompt, $[sep]$ is the separator and $y_i \in \{0, 1\}$ as shown in Figure 1 *(bottom)*. Label-prompt represents the ground-truth label of the sentence in textual format. For example, *this post contains hatespeech* / *this post contains normal words* can be used as label-prompt for *hate* and *non-hate* sentences respectively (Table B). The target to the model, $y_i = 0$ indicates that the sentence, $s_i$ and label-prompt, $l_j$ do not entail each other. $y_i = 1$ indicates entailment. We extend analysis of `Entailment-style` for multiple languages using monolingual and multilingual label-prompts.

## 2.3 Explicit and Implicit Hate Speech

In this section, we study the correlation between *explicit* and *implicit*. As discussed previously, *explicit* hate speech comprises of *cuss, swear, abusive, profane* words but *implicit* hate speech is more contextual and indirect. While the manner of expression is different, the intent behind both these modes is similar. To leverage this, we pretrain on the task of *explicit* hate speech detection and finetune it on *implicit* hate speech dataset and vice-versa and observe consistent gains. We probe the

| Model | Acc | F1 | ΔF1 |
|---|---|---|---|
| RoBERTa-base | 68.61 | 67.20 | - |
| RoBERTa-Tw | 69.18 | 67.64 | +0.44 |
| RoBERTa-TwS | 69.54 | 67.88 | +0.24 |
| RoBERTa-TwS-EasyMixup | **69.80** | **68.33** | +0.45 |
| Mathew et al. (2021) | 69.00 | 67.40 | - |

Table 1: Explicit Hate: Accuracy and F1 score on HateXplain dataset averaged over 3 runs.

| Model | Acc | F1 | ΔF1 |
|---|---|---|---|
| RoBERTa-base | 76.91 | 74.09 | - |
| RoBERTa-Tw | 77.86 | 75.77 | +0.68 |
| RoBERTa-TwS | 78.36 | 76.13 | +0.36 |
| RoBERTa-TwS-EasyMixup | **78.38** | **76.66** | +0.53 |
| ElSherief et al. (2021) | 77.50 | 70.40 | - |

Table 2: Implicit Hate: Accuracy and F1 score on LatentHatred dataset averaged over 3 runs.

layers of contextual models by extracting the features from each layer and training a classifier over these representations to understand how contextual models encode the information about hate speech and observe that *explicit* and *implicit* hate speech is encoded differently.

## 3 Dataset and Models

**Explicit:** We experiment with HateXplain (HX)(Mathew et al., 2021) dataset for *explicit* hate speech study. HateXplain (HX) captures explicit lexicon based hate speech posts collected from popular social media sites like Twitter and Gab.

**Implicit:** For *implicit* hate speech, we use LatentHatred (LH)(ElSherief et al., 2021), which comprises of *implicit* hate speech containing indirect/coded language.

**Multilingual:** We also experiment with *explicit* hate speech datasets in French (FR), Spanish (ES), Arabic (AR) and Portuguese (PT) [2] for evaluating our methodolgy for different languages. Since the taxonomy was different for each label, we focus on the datapoints annotated with *hate* and *non-hate* labels only (Poletto et al., 2021). In Appendix Section A, we summarize the details and statistics

[2]hatespeechdata.com

| Model | Accuracy | F1 Score | ΔF1 |
|---|---|---|---|
| RoBERTa-Tw | 69.18 | 67.64 | - |
| RoBERTa-Tw-IH | **70.74** | **68.88** | +1.24 |
| RoBERTa-Tw | 77.86 | 75.77 | - |
| RoBERTa-Tw-EH | **78.38** | **75.95** | +0.18 |

Table 3: Cross-Learning results between *explicit* and *implicit* hate speech detection.

| Lang | DL | XLM-R | XLM-Tw | XLM-TwS | EM-mo | EM-mu |
|---|---|---|---|---|---|---|
| FR | 65.95 | 64.48 | 68.36 | 72.73 | 78.58 | **81.16** |
| ES | 73.29 | 76.99 | 77.27 | 77.87 | 79.23 | **80.66** |
| AR | 83.20 | 82.36 | 83.57 | 84.50 | 84.80 | **85.60** |
| PT | 69.41 | 71.83 | 72.35 | 72.76 | 73.60 | **74.09** |

Table 4: F1 score on two-way classification (hate, non-hate) for different languages using adaptation and monolingual (EM-mo) and multilingual (EM-mu) variations of EasyMixup augmentation. DL((Aluru et al., 2020))

| | Baseline | | + prompt-en | | + prompt | |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 |
| HX | 69.85 | 68.36 | 72.97 | 71.39 | **72.97** | **71.39** |
| LH | 77.81 | 74.42 | 78.57 | 75.97 | **78.57** | **75.97** |
| FR | 88.46 | 84.62 | 88.55 | 84.64 | **94.23** | **92.83** |
| ES | 76.13 | 75.87 | 77.06 | 76.74 | **80.44** | **79.97** |
| AR | 89.67 | 78.09 | 89.30 | 78.51 | **90.41** | **82.03** |
| PT | 72.19 | 66.50 | 75.00 | 67.98 | **79.23** | **71.04** |

Table 5: F1 score on entailment task for all datasets using english prompts (prompt-en) and language-specific prompts (prompt). Baseline corresponds to BERT-base for HX, LH and mBERT for rest. For English datasets, prompt is equivalent to prompt-en.

of all the datasets.

**Models:** We consider RoBERTa-base (Liu et al., 2019) and XLM-R (Conneau et al., 2020) as the baseline model for English and other languages respectively. For exploring the impact of domain adaptive models, we experiment with RoBERTa-Tw and XLM-Tw models. For the multilingual experiments, we use XLM-TwS, which is the XLM-Tw model finetuned on the UMSAB dataset (Barbieri et al., 2021).More details in Appendix Section C.

## 4 Results

**Explicit:** In Table 1, we report the results on HateXplain dataset. We observe that RoBERTa-Tw improves upon the results of RoBERTa-base model. This shows that the pretraining over similar domain (social media) helps in achieving better performance. RoBERTa-TwS which has been trained for sentiment detection demonstrates further improvement highlighting the correlation between sentiments and hate-speech detection. On adding our augmentation (RoBERTa-TwS-EasyMixup), we notice further performance gains demonstrating the benefits of EasyMixup augmentation. Overall, our results improve upon the previously reported baseline (Mathew et al., 2021).

**Implicit:** We conduct similar experiments on LatentHatred dataset. We notice gains by using the domain adapted RoBERTa-Tw model.

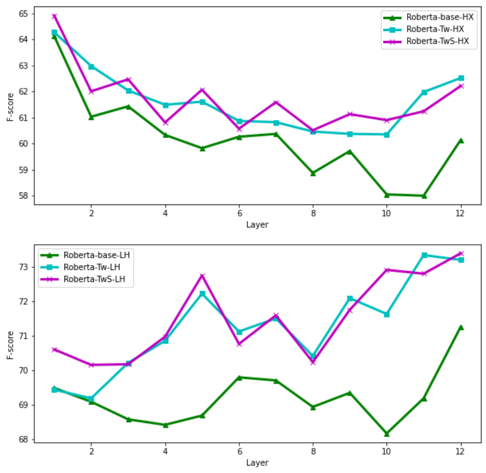


Figure 2: Layer-wise probing results on `HateXplain` (top) and `LatentHatred` (bottom) datasets for `RoBERTa-base`, `RoBERTa-Tw` and `RoBERTa-TwS` [Best viewed in color].

`RoBERTa-TwS` does not improve the accuracy but improves upon the F1 score which is a better metrics due to data imbalance. Addition of `EasyMixup` (`RoBERTa-TwS-EasyMixup`) further improves the performance. Our results improve upon the previously reported state-of-the-art results 2.

**Explicit-Multilingual:** We evaluate our method on 4 more languages in Table 4 and observe similar trends. For all the languages, multilingual domain (`XLM-Tw`) and task adapted (`XLM-TwS`) models perform better than the base model (`XLM-R`). On integration of `EasyMixup`, we further note improvements. We also experiment with sampling augmented samples from other languages (`EM-mu`) and notice further gains highlighting the cross learning between languages by 1-3%. We compare `EasyMixup` with state-of-the-art method `SSMixup` in Table 6 and observe that `EasyMixup` improves the performance by 1-2% for both implicit and explicit hate-speech detection.

**Entailment-style:** In Table 5, we report the results using monolingual[3] and English prompts and observe that monolingual prompts outperform English prompts. This is not surprising considering that models are trained on pairs of sentences from same language only. We use `mBERT`/ `BERT-base` for this study as it has been trained with NSP task which aligns with `Entailment-style`. Check Appendix B for more details.

**Implicit-Explicit Correlation:** We finetune the `RoBERTa-Tw` model on *implicit* hate speech (`RoBERTa-Tw-IH`) before training it for implicit hate speech and observe the F1 improvement from 67.64 to 68.88 in Table 3. This shows that *implicit* hate speech detection benefits the task of *explicit* hate speech. Similarly, F1 score of *implicit* hate speech detection improves from 75.77 to 75.95 by finetuning using *explicit* hate speech dataset.

[3]We used Google Translate to obtain monolingual prompt

| Model | Acc | F1 |
|---|---|---|
| **`LatentHatred`** | | |
| `BERT-base` | 76.51 | 73.70 |
| +`SSMixup` (Yoon et al., 2021) | 77.30 | 74.76 |
| +`EasyMixup` | **77.52** | **75.28** |
| **`HateXplain`** | | |
| `BERT-base` (Mathew et al., 2021) | 69.00 | 67.40 |
| +`SSMixup` | 69.59 | 67.72 |
| +`EasyMixup` | **69.70** | **68.66** |

Table 6: Comparing `EasyMixup` with SSMixup (Yoon et al., 2021)

**Probing:** In Figure 2, we plot the F1 score of `RoBERTa-base` and `RoBERTa-Tw` for *explicit* and *implicit* hate speech across different layers of the contextual model. We note that lower layers show higher F1 for *explicit* hate speech detection (expected layer = 0.98), while higher layers demonstrate better *implicit* hate detection performance (expected layer = 5.12). This alludes to the hypothesis that *implicit* hate speech is contextual in nature while *explicit* hate speech can be detected by using token-level information also. Training details are described in Appendix Section D.

## 5 Conclusion

In this work, we introduced a novel input-level data-augmentation technique, `EasyMixup` which shows performance gains over monolingual and multilingual settings. We also explored reformulation of hate speech classification as `Entailment-style` problem and achieved substantial performance gains using monolingual entailment. We also performed layer probing to find that higher layers encode *implicit* hate information, while lower layers are more focused on *explicit* hate speech highlighting the contextual nature of *implicit* and token-level dependence of *explicit* hate speech. In future work, we would like to explore how `EasyMixup` and `Entailment-style` perform when ensembled together in both mono, multi-lingual settings.

## 6 Limitations

One limitation would be that `EasyMixup` won't be applicable in tasks like sentiment analysis where the final mixed label might not be binary.

## References


Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. In *ECML-PKDD*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis Espinosa Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Findings of EMNLP*.

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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*, pages 8440–8451.

T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. *In Proceedings of the Eleventh International Conference on Web and Social Media, 512–515. Montreal, Qu ´ ebec, Canada: AAAI Press.*

Parag Dutta, Souvic Chakraborty, Sumegh Roychowdhury, and Animesh Mukherjee. 2022. Crush: Contextually regularized and user anchored self-supervised hate speech detection. *Findings of NAACL.*

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *EMNLP*.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3)*.

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. *ICWSM*.

Ona Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. *WACV*.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. Adima: Abuse detection in multilingual audio. *arXiv preprint arXiv:2202.07991*.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. 2020. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Abusive Language Online Workshop, EMNLP 2018*, pages 26–32.

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 preprint arxiv.1907.11692*.

B. Mathew, R. Dutt, P. Goyal, , and A Mukherjee. 2019. Spread of hate speech in online social media. *In Proceedings of the 10th ACM Conference on Web Science, 173–182.*

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *AAAI*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. *CoRR*, abs/1910.12574.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y Yeung. 2019a. Multilingual and multi-aspect hate speech analysis. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4667–4676.*

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019b. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systemic review. *Language Resources and Evaluation volume 55, pages477–523*.

Lara Quijano-Sanchez, Juan Carlos Pereira Kohatsu, Federico Liberatore, and Miguel Camacho-Collados. 2019. Haternet a system for detecting and analyzing hate speech in twitter. In *Zenodo*.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

F M Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. 2021. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *ICLR*.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML*.

Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. 2020. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *arXiv:2003.13048*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. Ssmix: Saliency-based span mixup for text classification. In *Findings of ACL*.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*.

Ziqi Zhang, D. Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *Extended Semantic Web Conference, ESWC 2018*.

# A Dataset

In Table 7, we note the dataset size and source of the datasets used in our study. Majority of the datasets are source from Twitter and have data imbalance.

**Explicit Hate (HX):** `HateXplain` dataset has been sourced from Twitter and Gab. The lexicon set from (Davidson et al., 2017), (Ousidhoum et al., 2019a) & (Mathew et al., 2019) is combined to sample 1% tweets in the period Jan-2019 to Jun-2020. For Gab, they use the dataset provided by (Mathew et al., 2019). All posts containing embedded links, pictures, videos were removed and usernames were anonymized by replacing with *user* token. Each post in the dataset is labelled into 3 categories: *Normal*, *Offensive* or *Hateful*. For the annotation task, Amazon Mechanical Turk (MTurk) workers are used where each post is labelled by 3 annotators and the ground truth class is chosen by majority voting. Finally, 19,229 posts were annotated of which 5,935 were hateful, 5,480 were offensive and 7,814 were normal. For the rest 919 posts the annotators provided 3 different classes and hence these were discarded.

**Implicit Hate (LH):** `LatentHatred` introduces a theoretically-justified taxonomy of implicit hatespeech with fine-grained labels on eight ideological clusters of US hate groups as given by the SPCL report - *Black Separatist*, *White Nationalist*, *Neo-Nazi*, *Anti-Muslim*, *Racist Skinhead*, *Ku Klux Clan*, *Anti-LGBT* and *Anti-Immigrant*. For high-level categorization, the tweets were categorized into *explicit hate*, *implicit hate* & *non-hateful*. Overall, the dataset contains 21,480 tweets, where 7,100 were implicit hate, 1,089 explicit hate and 13,291 non-hateful. Using majority vote, labels were obtained for 19,112 tweets of which 4,909 were implicit hate, 13,291 non-hateful and rest 933 explicit hate were discarded. For a finer categorization, 6 labels were chosen representing principal axes of implicit hate - *White Grievance*, *Incitement*, *Inferiority*, *Irony*, *Stereotypes* & *Threatening*. The 4,909 implicit hate tweets labeled in the high-level stage were further annotated using the above mentioned fine-grained labels.

**Multilingual:** We collected 6 publicly available datasets in 4 different languages - French, Spanish, Arabic and Portuguese and combined them individually. Each dataset had a variety of labels - *hate, abusive, profanity, offensive* etc. Since the taxonomy is different for each label, we focus on the datapoints annotated with *hate* and *non-hate* labels. We describe each dataset in following section.

| Dataset | Source | #datapoints | %hate |
|---|---|---|---|
| HateXplain | Twitter, Gab | 19,229 | 30.86 |
| LatentHatred | Twitter | 20,391 | 34.82 |
| Arabic | Twitter | 5,418 | 17.07 |
| Portuguese | Twitter | 5,670 | 31.53 |
| Spanish | Twitter | 11,150 | 33.29 |
| French | Twitter | 1,028 | 20.14 |

Table 7: Dataset Statistics

- **Arabic (AR)**: Mulki et al. (2019) contains Syrian/Lebanese political tweets labeled as abusive, normal or hate. (Ousidhoum et al., 2019b) consists of multi-labeled tweets based on attributes like hostility, target, directness, etc.
- **Spanish (ES)**: Basile et al. (2019) provided a multilingual hatespeech dataset against women & immigrants. Quijano-Sanchez et al. (2019) collected a small hatespeech dataset in spanish with hate/non-hate labels.
- **Portuguese (PT)**: Fortuna et al. (2019) provided a hierarchically labeled hatespeech dataset of which we use only the binary labels for our task.
- **French (FR)**: Ousidhoum et al. (2019b) consists of multi-labeled tweets based on attributes like hostility, target, directness, etc.

# B Prompts used for `Entailment-style` task

Refer to Table 8.

# C Model Details

`RoBERTa-Tw` is based on `RoBERTa-base` model trained on 60M English tweets. `XLM-Tw` (Barbieri et al., 2021) is a `XLM-R` model trained on 200M tweets retrieved from 30+ languages. For task-adaptive models, we take `RoBERTa-TwS` and `RoBERTa-Tw-EH` which are initialized with the `RoBERTa-Tw` model and further finetuned using Sentiment and Hatespeech classification data from the TweetEval (Barbieri et al., 2020) benchmark.

# D Implementation Details

We perform all experiments with 3 different seeds on a single NVIDIA V100 GPU and report the

| Language | Label Description |
|---|---|
| HateXplain | *this post contains hate speech / this post contains {offensive,normal} words* |
| LatentHatred | *this is implicit hate / this is normal* |
| French | *c'est odieux / c'est normal* |
| Spanish | *esto es odioso / esto es normal* |
| Arabic | هذا المنشور يحتوي على كلام يحض على الكراهية / هذا المنشور يحتوي على كلمات عادية |
| Portuguese | *este post contém discurso de ódio / este post contém palavras normais* |

Table 8: Prompts used across various datasets for `Entailment-style` task.

average score. We use a batch size of 16 and maximum sequence length of 128. We choose initial learning rate from {3e-5, 4e-5, 5e-5} and perform linear decay after 10% warmup steps. We use the AdamW optimizer and train our models for 5 epochs. The classifier head consists of a 2-layer MLP with ReLU activation. We choose the best checkpoint using validation metrics every epoch. From our experiments, we found best reported results were obtained by combining *offensive+normal* & *hate+normal* classes for HateXplain and *hate+normal* classes for LatentHatred and keeping $p_{aug} = 0.2$ and $p_{flip} = 0.5$.

For the probing experiments, we train the 2-layer MLP probe classifier for 50 epochs with batch size 64 and learning rate 1e-3.

For the entailment experiments, we use a batch size 128 (required for entailment method to get good gains) consistently for all methods and learning rate 3e-5.

## E Effect of Length

We used the max sequence length of 128 in our experiments. $< 1\%$ of samples exceed this limit across all datasets - `HateXplain`, `LatentHatred`, `MultilingualHate`. Thus, length of 128 tokens does not degrade `Entailment-style` performance. However, in case of `EasyMixup`, length of concatenated sentences could exceed 128 tokens. To evaluate the impact, we repeat experiments using best performing model - `RoBERTa-TwS-EasyMixup` (averaged over 3 random seeds) keeping maximum sequence length as 512. For `HateXplain`, $\Delta$ Accuracy / F1 $\sim$ 0.00 / -0.03 % and for `LatentHatred` $\Delta$ Accuracy / F1 $\sim$ +0.21 / -0.05 %. As we can see there is no significant impact from the reported results. This can be attributed to the fact that we do probabilistic mixup in `EasyMixup` ($p_{aug} = 0.2$ and $p_{flip} = 0.5$). Thus the model sees all type of examples during the training phase.

## F Ethical Considerations

All the datasets that we use are publicly available. We report only aggregated results in the main paper. We have not or do not intend to share any Personally Identifiable Data with this paper. We release the code and data associated with this paper as well - https://anonymous.4open.science/r/data_efficient_hatedetect/