Respectful or Toxic?
Using Zero-Shot Learning with Language Models to Detect Hate Speech

Flor Miriam Plaza-del-Arco, Debora Nozza, Dirk Hovy
Bocconi University
Via Sarfatti 25
Milan, Italy
{flor.plaza, debora.nozza, dirk.hovy}@unibocconi.it

Abstract

Hate speech detection faces two significant challenges: 1) the limited availability of labeled data and 2) the high variability of hate speech across different contexts and languages. Prompting brings a ray of hope to these challenges. It allows injecting a model with task-specific knowledge without relying on labeled data. This paper explores zero-shot learning with prompting for hate speech detection. We investigate how well zero-shot learning can detect hate speech in 3 languages with limited labeled data. We experiment with various large language models and verbalizers on 8 benchmark datasets. Our findings highlight the impact of prompt selection on the results. They also suggest that prompting, specifically with recent large language models, can achieve performance comparable to and surpass fine-tuned models, making it a promising alternative for under-resourced languages. Our findings highlight the potential of prompting for hate speech detection and show how both the prompt and the model have a significant impact on achieving more accurate predictions in this task.

1 Introduction

The rising prevalence of online hate speech and its harmful effects have made hate speech detection a central task in natural language processing (NLP). Despite progress, the prevalent supervised learning approaches encounter significant challenges: many languages or contexts have little or no labeled data (Poletto et al., 2021). Hate speech is also subjective and context-dependent, as it is influenced by factors such as demographics, social norms, and cultural backgrounds (Talat and Hovy, 2016).

To overcome these challenges, approaches like zero-shot learning (ZSL) and prompting of large language models (LLMs) have emerged.[1] Both use a *template* to process the original text and the class labels as *verbalizers*. This approach leverages the LLM's knowledge to predict the likelihood of the (class) verbalizers in the template. These verbalizers guide the model's understanding of a specific task. For binary hate speech detection, the template might be "`<text>`. This text is `<verbalizer>`", where `<verbalizer>` can be "`hateful`" or "`non-hateful`". For the input, "I hate you. This text is", the LLM should associate a higher likelihood with the verbalizer completion "`hateful`". By picking the more likely completion, this approach requires no training data. It has shown promising results in various NLP applications (Zhao et al., 2023; Su et al., 2022; Wei et al., 2022; Brown et al., 2020).However, to date, its effectiveness for hate speech detection remains largely unexplored.

We comprehensively evaluate ZSL with prompting for hate speech detection to better understand its capabilities. The choice of appropriate verbalizers is a key factor in the effectiveness of prompting (Plaza-del-Arco et al., 2022; Liu et al., 2023). To this end, we systematically compare various verbalizers across multiple models. We evaluate the performance of conventional transformer models and more recent instruction fine-tuned LLMs on 8 benchmark datasets to assess their robustness. Furthermore, we test our approach on two languages with limited labeled data (Italian and Spanish). Our results show that ZSL with prompting matches or surpasses the performance of fine-tuned models, particularly in instruction fine-tuned models.

Contributions 1) We investigate the effectiveness of ZSL with prompting for hate speech detection 2) We conduct a systematic exploration and comparison of various verbalizers across 5 models 3) We extend our investigation to two languages with limited labeled data. Our code is publicly available at `https://github.com/MilaNLProc/prompting_hate_speech`.

[1]Note that ZSL could be used with various models, whereas prompting is specific to LLMs. Here, we use ZSL to prompt LLMs without additional labeled examples in the prompt (few-shot learning), but only the target sentence.

The 7th Workshop on Online Abuse and Harms (WOAH), pages 60–68

2 Datasets

We compare our results on 8 benchmark datasets using binary classification. See Table 1 for details. They differ in terms of size, corpus source, and labels. More details are in Appendix A.

Dataset	Size	Source
DAVIDSON	24,802	Twitter
DYNABENCH	41,255	Synthetic
GHC	27,665	Gab
HATEVAL	13,000	Twitter
HATEXPLAIN	20,148	Twitter and Gab
MHS	50,000	Youtube, Twitter and Reddit
MLMA	5,647	Twitter
HSHP	16,914	Twitter

Table 1: Datasets used in our experiments.

3 Prompting for Zero-Shot Hate Speech Classification

We use ZSL with prompting to evaluate the models' ability to detect hate speech. First, we test various encoder models to select the best verbalizers. We then test those verbalizers on recent instruction fine-tuned LLMs and compare to encoder models.

Encoder-based Language Models For our experiments, we use the following prompt template: "<text> This text is <verbalizers>". We then check the LLM likelihood of hateful and non-hateful verbalizers and select the most probable completion as final prediction. We test all 25 possible pairs from the following lists. For hate: `harmful`, `abusive`, `offensive`, `hateful`, `toxic`, and for non-hate `respectful`, `kind`, `polite`, `neutral`, `positive`.

We compare three different language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020). We use OpenPrompt (Ding et al., 2022), a standard framework for prompt learning over pre-trained language models.

Instruction Fine-tuned Language Models We experiment with recent instruction fine-tuned language models. They are fine-tuned on a large set of varied instructions that use an intuitive description of the downstream task to answer natural language prompts. In this approach, we formulate the prompt template as "Classify this text as <$\text{verb}_{\text{non-hate}}$> or <$\text{verb}_{\text{hate}}$>. <text>. Answer:", for the *verbalizers* ($\text{verb}_{\text{non-hate}}$, $\text{verb}_{\text{hate}}$) we consider the best pair obtained with the encoder models, and for the prompt models, we use the Fine-tuned Language Net (FLAN-T5) model (Chung et al., 2022) and mT0 (Muennighoff et al., 2022). Note that FLAN-T5 has been trained for toxic language detection.

Baseline We used (1) a RoBERTa model fine-tuned with supervised training on each hate speech dataset and (2) the commercial Perspective API.[2]

4 Results

4.1 Encoder models

Table 2 shows the results of several encoder models on multiple hate speech detection benchmark datasets. Overall, the best-performing encoder model across different datasets is $\text{RoBERTa}_{\text{LARGE}}$ obtained the best macro-F1 score in 5 out of 8 datasets. Regarding the verbalizers, the pair `positive` and `polite` yield the best results in identifying non-hateful speech, while `hateful` and `toxic` prove best for detecting hate speech. This highlights the need for careful selection of verbalizers to achieve optimal performance in this task.

Identifying Best Verbalizers To select the best pair of verbalizers that work well across models and datasets for hate speech detection, we averaged the different performance metrics by model and dataset across all folds. As shown in Table 3, the best-performing verbalizer pair is `respectful-toxic`, which achieves the highest macro-F1 score of 42.74. The verbalizers most commonly associated with the non-hate speech class are `respectful` and `polite`, while `toxic` and `hateful` are more commonly associated with hate speech. We select the best verbalizer pair (`respectful-toxic`) to conduct additional experiments.

4.2 Encoder vs. Instruction Fine-tuned LLMs

In this section, we compare the results obtained by prompting the encoder-based models and the instruction fine-tuned models. The results are shown in Table 4. These models are prompted using the best pair of verbalizers we found in the encoder-based models, which is `respectful-toxic`. In general, the recent models mT0 and FLAN-T5 outperform the encoder-based models by a large margin showing an average improvement of 39.75% and 65.33% over the encoder models, respectively. In particular, FLAN-T5 exhibits remarkable performance in detecting hate speech across various datasets, which can be attributed to its prior fine-tuning for toxic detection. This suggests that the

[2] https://www.perspectiveapi.com/

Dataset	Model	$\text{Verb}_{\text{non-hate}}$	$\text{Verb}_{\text{hate}}$	$\text{F1}_{\text{non-hate}}$	F1_{hate}	Macro-F1
DAVIDSON	$\text{RoBERTa}_{\text{LARGE}}$	positive	hateful	41.38	69.15	55.26
DYNABENCH	$\text{RoBERTa}_{\text{LARGE}}$	positive	harmful	52.96	57.36	55.16
GHC	$\text{RoBERTa}_{\text{LARGE}}$	positive	hateful	45.03	68.85	56.94
HATEVAL	$\text{BERT}_{\text{BASE-uncased}}$	polite	toxic	61.52	58.05	59.78
HATEXPLAIN	$\text{RoBERTa}_{\text{LARGE}}$	polite	toxic	24.36	86.23	55.30
MHS	$\text{RoBERTa}_{\text{LARGE}}$	positive	hateful	66.91	73.68	70.30
MLMA	$\text{DeBERTa}_{\text{V3-BASE}}$	polite	hateful	12.32	93.53	52.93
HSHP	$\text{RoBERTa}_{\text{BASE}}$	positive	hateful	73.79	54.64	64.21

Table 2: Class and macro-F1 score of encoder models on different benchmark datasets.

Verb-nh	Verb-h	F1-nh	F1-h	Macro-F1
respectful	toxic	27.28	58.19	42.74
polite	hateful	24.37	59.42	41.89
positive	hateful	34.58	48.84	41.71
positive	offensive	19.37	63.94	41.66
neutral	toxic	31.17	52.11	41.64
respectful	hateful	18.60	63.91	41.25
polite	toxic	28.30	53.79	41.04

Table 3: Verbalizer pairs across encoder models and datasets by Macro-F1 score.

knowledge learned from detecting toxic language is transferable and can be leveraged to improve hate speech detection in other datasets. In addition, we conduct a comparison between the supervised learning upper bound, a fine-tuned $\text{RoBERTa}_{\text{BASE}}$ model, and the instruction fine-tuned models in our ZSL experiments. Our findings show that the instruction fine-tuned models achieve comparable performance, and FLAN-T5 even surpasses the $\text{RoBERTa}_{\text{BASE}}$ fine-tuned model in some datasets, such as GHC, HATEXPLAIN, and MLMA. Overall, the DAVIDSON dataset achieves the highest performance among all the datasets, with a macro-F1 score of 83.30. In contrast, the MLMA dataset obtains the lowest macro-F1 score of 54.35, which is expected given its complexity arising from the low inter-annotator agreement. Notably, the performance on the HATEVAL dataset (65.38) exhibits an improvement over the participant results' mean (44.84) in the competition (Basile et al., 2019). On the DYNABENCH dataset, the FLAN-T5 model's result (58.08) is similar to that of fine-tuning the $\text{RoBERTa}_{\text{BASE}}$ fine-tuned model (61.76), despite the dataset's complexity with a large number of challenging perturbations that make it harder for models to detect hate speech accurately. Finally, we compared our approach with Perspective API, the most popular commercial tool for toxicity detection. FLAN-T5 is outperforming it in 6 cases out of 8, demonstrating prompting to be a more accurate solution. While the varying degrees of difficulty across datasets in hate speech detection is demonstrated in these results, the potential of instruction fine-tuned models to achieve state-of-the-art performance on various benchmarks without requiring fine-tuning on a specific dataset is highlighted. This insight is especially valuable for subjective tasks like hate speech, where the complex nature of labeling this phenomenon can make it challenging to find labeled datasets.

5 Results on Multi-Lingual Datasets

We also investigated the effectiveness of ZSL with prompting in a multilingual context, which is often more challenging due to the scarcity or unavailability of data. We present the outcomes achieved by multilingual models: multilingual XLM-R (Conneau et al., 2020) as encoder model and mT0 and FLAN-T5 as instruction fined-tuned models. The prompt has been written in English following the same templates presented in Section 3 and using the best-performing verbalizer pair `respectful-toxic`. We use the experimental settings adopted in Nozza (2021), comparing our method with their fine-tuned XLM-R model. Thus, the dataset comprises English (EN), Spanish (ES), and Italian (IT). The HatEval (Basile et al., 2019) shared task dataset on hate speech against immigrants and women on Twitter is adopted for English and Spanish. For Italian, two different corpora proposed for Evalita shared tasks (Caselli et al., 2018) are considered: the automatic misogyny identification challenge (AMI) (Fersini et al., 2018) for hate speech towards women, and the hate speech detection shared task on Facebook and Twitter (HaSpeeDe) (Bosco et al., 2018) for hate speech towards immigrants.

The results are shown in Table 5. Regarding the ZSL approaches, the instruction fine-tuned models outperform XLM-R, with FLAN-T5 achieving the highest macro-F1 score on all languages. The

Dataset	ZSL Prompting							API	Fine-tuning
	RoBERTa$_B$	RoBERTa$_L$	BERT$_B$	DeBERTa$_B$	DeBERTa$_L$	mT0	FLAN-T5	Perspective API	RoBERTa$_B$
DAVIDSON	42.46	40.87	52.33	46.67	25.99	54.46	83.30	79.20	**91.28**
DYNABENCH	36.68	36.08	45.87	51.38	37.57	54.11	58.08	55.50	**61.76**
GHC	42.02	41.43	53.13	50.13	35.36	56.07	61.53	**62.35**	59.59
HATEVAL	31.89	29.90	59.69	55.82	36.68	57.76	65.38	60.77	**70.98**
HATEXPLAIN	49.38	46.11	48.93	51.67	20.88	56.68	**67.11**	58.86	60.34
MHS	44.60	36.16	62.23	57.38	43.29	74.70	79.38	87.90	**90.50**
MLMA	47.90	47.65	49.47	49.10	28.23	44.97	**54.35**	43.91	47.47
HSHP	27.50	24.77	43.10	44.17	40.37	53.97	64.36	56.30	**76.82**
Avg.% ↑	—	—	—	—	—	39.75 ↑	65.33 ↑	—	—

Table 4: Macro-F1 scores for different models on benchmark datasets using *respectful-toxic* verbalizer. B = base model, L = large model. Best model in bold, second-best underlined. Last row shows the average improvement of Flan-T5 and mT0 over encoder models.

Lang	XLM-R	mT0	FLAN-T5	Nozza (2021)
EN	29.80	57.85	**65.34**	41.6
ES	29.42	53.75	62.61	**75.2**
IT	31.34	43.25	57.29	**80.4**

Table 5: Macro-F1 scores on different languages. Best model in bold, second-best underlined.

ZSL models, as expected, did not outperform the fine-tuned XLM-R. However, the results obtained from the ZSL models are still considered adequate. Spanish, in particular, achieves comparable results with FLAN-T5 to the fine-tuned XLM-R. FLAN-T5 achieves better results in English because it is not affected by overfitting issues that arise during training (Nozza, 2021). These findings suggest that prompting with instruction fine-tuned LLMs is a promising method for hate speech detection in both mono and multilingual settings, without language-specific fine-tuning.

6 Related Work

Hate speech classification received increased attention in recent years. Supervised learning methods are the most common (Poletto et al., 2021; Fortuna et al., 2022). Among these methods, fine-tuning transformer-based LLMs emerged as the dominant paradigm (Plaza-del-Arco et al., 2020; Sarkar et al., 2021; Singh and Li, 2021; Caselli et al., 2021; Kirk et al., 2022, inter alia). However, they face significant challenges, like the limited availability of labeled data, especially in languages other than English (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), and the subjective nature of hate speech, which varies based on cultural background, personal experiences, and individual beliefs.

LLMs have led to innovative techniques like prompting (Liu et al., 2023) that use zero-shot and few-shot learning paradigms without needing labeled data. Recent works have explored these new techniques for hate speech detection. Chiu et al. (2021) use the prompts "Is the following text sexist? Answer yes or no" and "Classify the following texts into racist, sexist, or neither" to detect hate speech, with GPT-3 showing that LLMs have a role to play in hate speech detection. Schick et al. (2021) explore toxicity in LLMs using comparable prompts to self-diagnose toxicity during the decoding. They use the RealToxicityPrompts dataset (Gehman et al., 2020). (Goldzycher and Schneider, 2022) develop NLI-based zero-shot hate speech detection approaches using prompts as a hypothesis as proposed by Yin et al. (2019). Their results outperform fine-tuned models. Our work ZSL for hate speech classification differs from previous approaches as follows. (1) We provide a comprehensive evaluation of ZSL with prompting on multiple benchmark datasets, offering new insights into the effectiveness of this technique. (2) We explore the impact of the selection of verbalizers and models for the task, and (3) we compare the performance of encoder models with the recent LLMs based on instruction fine-tuning.

7 Conclusion

This paper presents a comprehensive evaluation of ZSL with prompting for hate speech classification. We have compared both encoder and instruction fine-tuned LLMs. Our experiments across different benchmark data sets showed that ZSL with prompting is a promising option to address the challenges presented in supervised learning systems. However, it also highlights the importance of carefully selecting the model and appropriate verbalizers, as they can significantly affect performance. Our results also show that recent LLMs based on instruction

fine-tuning play an essential role in hate speech detection. Further exploration of prompt formulation could lead to their continued growth in this area. Additionally, our multilingual experiments show that our proposed methods can be applied to other languages with comparable results.

Future research could investigate the bias presence (Dixon et al., 2018; Attanasio et al., 2022) and robustness (Röttger et al., 2021, 2022) of ZSL prompting for hate speech detection models, also in multilingual settings.

Limitations

While promising, our work presents limitations that need to be acknowledged. Firstly, we did not explore the best verbalizers for instruction fine-tuned language models, which could have further enhanced the performance of the models explored in this study, due to computational cost and the specific goals of the research. Secondly, we selected benchmark datasets based on their popularity and diversity, which might not be representative of all possible datasets in hate speech detection. We also acknowledge that, in addition to the languages examined in this paper, there are a number of other languages that may present unique challenges and characteristics for detecting hate speech. Our decision as to which languages to include in the multilingual experiment was based on a direct comparison with state-of-the-art research. Finally, we utilized the latest open-source language models for our experiments, but we did not explore other recent language models, such as the GPT family, primarily because they are not open and reasonably reproducible[3], and therefore the community may encounter challenges in replicating our results. These limitations provide directions for future research to improve and expand upon our work.

Ethics Statement

To ensure data privacy and protection, we use publicly available benchmark datasets for hate speech detection and do not collect any personal or sensitive information. Additionally, we acknowledge that the detection of hate speech can be a sensitive topic; therefore, we report the results of our experiments in a responsible and appropriate manner. Lastly, we acknowledge that language models trained on large datasets have the potential to perpetuate bias and discrimination, and we strive to mitigate these risks by carefully selecting and evaluating our models and verbalizers to ensure fairness and impartiality.

[3] https://hackingsemantics.xyz/2023/closed-baselines/

Acknowledgements

This project has in part received funding from Fondazione Cariplo (grant No. 2020-4288, MONICA). The authors are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263, pages 1–9, Turin, Italy. CEUR.org.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting Hate Speech with GPT-3. *arXiv preprint arXiv:2103.12407*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *International Conference on Web and Social Media*.

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*.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, 12:59.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP Practices Applied to Online Hate Speech Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. 51(4).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Janis Goldzycher and Gerold Schneider. 2022. Hypothesis engineering for zero-shot hate speech detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 75–90, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Wang, Weizhu Li, Yelong Liu, and Jianfeng Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020a. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Hannah Kirk, Bertie Vidgen, and Scott Hale. 2022. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech 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 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Flor Miriam Plaza-del-Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Flor-Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña López, and M. Teresa Martín-Valdivia. 2020. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.*, 20(2).

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

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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)*, pages 41–58, Online. Association for Computational Linguistics.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

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*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sumer Singh and Sheng Li. 2021. Exploiting auxiliary data for offensive language detection with bidirectional transformers. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–5, Online. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

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

Bertie Vidgen, Tristan Thrush, Zeerak Talat, 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)*, pages 1667–1682, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Dataset Details

Vidgen et al. (2021) (DYNABENCH) introduced a novel framework for dynamically creating benchmark corpora. The task assigned to the annotators involved identifying adversarial examples, which are instances that would be classified incorrectly by the target model and are particularly challenging to detect. The dataset contains a significant proportion of hateful entries, accounting for 54% of the dataset.

Kennedy et al. (2020b) (MHS) gathered a large collection of comments from diverse social media platforms (YouTube, Twitter, and Reddit). To label the comments, they used a crowdsourcing platform where four different ratings were given to each comment. To ensure a comprehensive assessment, the authors made certain that every annotator evaluated comments that spanned the entire hate speech scale. Since the dataset is annotated with a continuous hate score, we used a threshold set to binarise the problem: if value $<$ -1 $\rightarrow$ 0 and if value $>$ 0.5 $\rightarrow$ 1.

Kennedy et al. (2022) (GHC) presented the Gab Hate Corpus, a multi-label English dataset of posts sourced from gab.com, a social networking platform. To label the comments, at least three annotators labeled them under one of the following categories: Call for Violence, Assault on Human Dignity, or Not Hateful. Following Kennedy et al. (2020a), we aggregate the first two for obtaining the hateful class.

Basile et al. (2019) (HATEVAL) created the HatEval corpus for the HatEval campaign in SemEval. The dataset consists of tweets that were manually annotated via crowdsourcing for hate speech. To collect the tweets, they follow three different strategies: (1) monitoring potential victims of hate accounts, (2) downloading the history of identified haters, and (3) filtering Twitter streams with keywords, i.e., words, hashtags, and stems. The corpus contains a total of 24,802 tweets.

Talat and Hovy (2016) (HSHP) provided a dataset consisting of 16,914 tweets that were collected using Twitter's streaming API and filtered using a set of hate speech-related keywords related to religious, sexual, gender, and ethnic minorities. The tweets were then manually annotated by two annotators for the presence of hate speech.

Davidson et al. (2017) (DAVIDSON) created a dataset of 24,802 tweets annotated for the presence of hate speech and offensive language. The tweets were crawled using keywords related to a hate speech lexicon. Each tweet was labeled by three or more people into one of three categories: hate speech, offensive language, or neither. We aggregate the first two for obtaining the hateful class.

Mathew et al. (2021) (HATEXPLAIN) collected English posts from Twitter and Gab social media platforms. Afterward, a crowdsourcing platform was employed to categorize each post into three categories: hate speech, offensive speech, or normal speech. In addition to this, the annotators were tasked with identifying the target communities mentioned in the posts, as well as the specific portions of the post which formed the basis of their labeling decision. Finally, the majority voting decision was used to determine the final label. By combining the hate and offensive targets, the hateful class was formed. We combine the hate and offensive posts to obtain the hateful class.

Ousidhoum et al. (2019) (MLMA) presented a multilingual multi-aspect hate speech dataset comprising English, French, and Arabic tweets that encompass various targets and hostility types. Each tweet is labeled by 5 annotators, and then the majority vote is used to decide the final label. The average Krippendorff scores for inter-annotator agreement (IAA) are 0.153, 0.244, and 0.202 for English, French, and Arabic, respectively.

B Implementation Details

We implement the fine-tuned version of RoBERTa$_{\text{BASE}}$ with the following hyperparameter configuration for training: epochs are set to 3, batch size to 8, and the number of epochs

to 3. For the ZSL models, we used the default hyperparameters presented in Hugging Face. We fine-tune RoBERTa$_{\text{BASE}}$ for three epochs. We perform 5-fold partitions and report the results on the test set.

Hugging Face model cars BERT$_{\text{BASE-uncased}}$[4], RoBERTa$_{\text{BASE}}$[5], RoBERTa$_{\text{LARGE}}$[6], DeBERTa$_{\text{V3-BASE}}$[7], DeBERTa$_{\text{V3-LARGE}}$[8], XLM-RoBERTa$_{\text{LARGE}}$[9], mT0[10], and FLAN-T5[11].

Computing Infrastructure We run the experiments on one machine with the following characteristics: it is equipped with three NVIDIA RTX A6000 and has 48GB of RAM.

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/roberta-base
[6] https://huggingface.co/roberta-large
[7] https://huggingface.co/microsoft/deberta-v3-base
[8] https://huggingface.co/microsoft/deberta-v3-large
[9] https://huggingface.co/xlm-roberta-large
[10] https://huggingface.co/bigscience/mT0-xxl
[11] https://huggingface.co/google/flan-t5-xl