HARALD: Augmenting Hate Speech Data Sets with Real Data

Tal Ilan
Department of Ind. Eng. Manag.
Ben-Gurion University of the Negev
talilan13@gmail.com

Dan Vilenchik
School of Comput. Electr. Eng.
Ben-Gurion University of the Negev
vilenchi@bgu.ac.il

Abstract

The successful completion of the hate speech detection task hinges upon the availability of rich and variable labeled data, which is hard to obtain. In this work, we present a new approach for data augmentation that uses as input real unlabelled data, which is carefully selected from online platforms where invited hate speech is abundant. We show that by harvesting and processing this data (in an automatic manner), one can augment existing manually-labeled datasets to improve the classification performance of hate speech classification models. We observed an improvement in F1-score ranging from 2.7% and up to 9.5%, depending on the task (in- or cross-domain) and the model used.

1 Introduction

Hate speech detection (offensive, abusive, toxic) is of interest to academic researchers in a variety of domains, including computer science (Machova et al., 2020) and sociology (Davidson et al., 2017). It is also of interest to online social platforms that wish to maintain certain standards of discourse or are obliged to do so by law in some countries.

Hate speech is commonly defined as "any communication that disparages a target group of people based on some characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby, 2000).

Detecting hate speech may be difficult because the manifestation of speech as hate speech depends on a non-trivial interaction between various circumstances such as the topic, the context, the timing, outside events, and the identity of the speaker and recipient (Schmidt and Wiegand, 2017).

The typical way hate speech detection is approached is as a supervised-learning classification task, where lexical features and other features (e.g. word-embedding) are used to train a classifier (Schmidt and Wiegand, 2017; Spertus, 1997; Razavi et al., 2010b).

To that end, access to labeled corpora is essential. Since hate speech has many facets and there are few (or none) universal "gold-standard" datasets, authors usually collect and label their data. The size of collected corpora varies considerably ranging from around 100 labeled comments (Dinakar et al., 2012) to several thousand (Van Hee et al., 2015; Djuric et al., 2015).

Collecting and annotating hate-speech data is challenging and extremely time-consuming for two main reasons. First, there are much fewer hateful than benign comments present in randomly sampled data. Second, to manually annotate a data set, either expert annotators or crowd-sourcing services, such as Amazon Mechanical Turk, are employed. While crowd-sourcing has obvious advantages for this task, the annotation quality of non-expert annotators was demonstrated to be poorer than that of experts (Nobata et al., 2016; Ross et al., 2017).

There were several attempts to deal with these problems. (Waseem and Hovy, 2016a) proposed to select the text to be annotated by looking for topics that are likely to contain a higher degree of hate speech. They collected 136,052 tweets, about 10% were annotated (16,914 tweets), and about a third of them were labeled as hate speech. While this increased the proportion of hate speech posts, it focuses the resulting data set on specific topics and thus hinders the generalizability to other domains (Wiegand et al., 2019).

Another possible solution to the aforementioned two challenges may be found in data augmentation; this avenue is intensively developed for example, in computer vision but "relatively under-explored" in NLP, where the generation of effective augmented examples is "less obvious" (Feng et al., 2021).

2 Our Contribution and Method

This paper presents a new data augmentation pipeline for offensive/hate speech data called HARALD, which stands for Hate Augmentation with

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2241–2248

ReAL Data. Unlike common data augmentation methods that generate synthetic data (using GANs or other generative models) HARALD outputs an endless stream of relevant real data written by a huge number of authors, rich with various stylistic, grammatical, and semantic forms. Our method hinges upon the existence of online platforms where people are explicitly asked to be abusive. One such platform is the subreddit `r/RoastMe/`, in which users upload a picture and ask their peers to "roast them", with the intention to develop a thicker skin by withstanding the abusive speech (the logo of the subreddit is "the thicker the skin, the better the roast "). See the appendix for an excerpt from RoastMe.

To validate HARALD's usefulness, we conducted the following experiment, inspired by the cross-domain evaluation of (Wiegand et al., 2019). We harvested from `r/RoastMe/` a total of 3700 messages and assigned each message a hate score, the output of the last GELU layer of a pre-trained BERT model for hate speech detection (Caselli et al., 2020). We then sorted the messages and took the top 1,000 as the positive class of the RoastMe (RM) dataset. We then selected six well-known datasets of hate/offensive speech to fine-tune the BERT-base-uncased model (Devlin et al., 2018) on each of the datasets separately (see Section 4 for details on the datasets). We tested the cross-domain performance of each of the six models on the other five datasets. We then repeated this experiment, but now we added another fine-tune step with the RM dataset. We also conducted an in-domain (cross-validation) test. All detail in Section 5.

We observed an improvement in macro F1-score ranging from 2.7% and up to 9.5%, depending on the task (in or cross domain) and on the model that was used (see Table 3). For the (Waseem and Hovy, 2016a) dataset, we obtained a 4.1% improvement when using RM in the in-domain task, improving the F1 score from 0.74 to 0.77. For comparison, the GAN-based pipeline of (Cao and Lee, 2020) improved the F1 score from 0.77 to 0.78 on the same dataset (1.2% improvement).

3 Related Work

Data augmentation methods have been explored to address the imbalance of datasets challenge in supervised classification tasks. Noise injection or attribute modification techniques were commonly applied to generate synthetic data for image and sound classification tasks (Shorten and Khoshgoftaar, 2019; Tran et al., 2017; Salamon and Bello, 2017). However, such techniques do not extend to text due to the categorical nature of words and the sequential nature of text.

There are very few works that explored data augmentation in hate speech detection. (Rizos et al., 2019), and similarly (Ibrahim et al., 2018), explored various data augmentation techniques for hate speech: substituting words, swapping word positions, and neural generation using RNN (Sutskever et al., 2011).

Each of these methods has its limitations. It is challenging to find suitable semantically similar words in the fast-evolving social media platforms; swapping words' positions may harm the coherence of the sentence.

The authors of (Cao and Lee, 2020) propose a GAN methodology, HateGAN, to augment two data sets. They train LSTM and CNN models on the augmented datasets and show a 5% improvement in F1 score. They also show that HateGAN outperforms (Rizos et al., 2019).

In (Dixon et al., 2018), real non-toxic text was harvested similarly to us, but for the task of mitigating unintended biases in text classification. One has to note, though, that most online text is non-toxic, so automatically harvesting toxic or non-toxic text is by no means equivalent tasks.

Our work differs from these works in several key aspects. (1) HARALD produces real rather than synthetic data, the distribution of which is different than the dataset to be augmented. Previous work generates synthetic data from the existing dataset and makes a point that the data has the same distribution as the data to be augmented. (2) We train SOTA hate speech classification models, BERT, while weaker models such as LSTM or CNN were used in previous work. (3) HARALD improves at a more challenging task – cross-domain prediction. We surmise that the fact that RM has a different distribution than the original dataset plays a key factor in improving the prediction results. (4) We evaluate HARALD in six different datasets, while previous work used a maximum of three.

Finally, let us discuss the subreddit `r/RoastMe`. RoastMe presents an intriguing case of how alternative norms can emerge in online communities, allowing behaviors that are otherwise condemned as inappropriate to be reframed as acceptable. In this community, users post photos

of themselves with the explicit expectation of being mocked or ridiculed by others. RoastMe is not alone, with similar subreddits such as `r/ToastMe`, and `r/Judgemeplease`. The norms and values of the RoastMe community were studied, for example, in (Kasunic and Kaufman, 2018; Allison et al., 2019). In (Sodhi et al., 2021), RoastMe was used for the task of style transfer, rephrasing slurs as compliments and vice versa.

4 Data

We turn to describe the RM dataset and the other six datasets that we augmented using RM in order to evaluate the performance of HARALD. All datasets appear in the project's GitHub page (Ilan and Vilenchik, 2022).

The RoastMe (RM) dataset. That paper overall supports our thesis (to quote, "r/RoastMe, a comedy-focused subreddit of the parent site reddit.com, wherein members post photos of themselves to be ridiculed by other members; the site generally encourages harsh and offensive forms of humor in these interpersonal exchanges").

We harvested 3700 comments from the Roastme using the PRAW API. We removed comments with less than three words, and cleaned them from links, emojis, stop words, and punctuation marks, leaving us with 3,500 comments. We then used the HateBERT from (Caselli et al., 2020), further fine-tuned on the Kaggle dataset (see below), to assign each RM comment a hate score (output of last GELU layer). We sorted the comments in descending order and took the top 1,000 as the positive class of the RM dataset.

The RoastMe dataset also contains a negative class to keep the train and test balanced after augmentation. We sampled 3,500 non-offensive Reddit comments from the (Qian et al., 2019) dataset (see below), ranked them using the same BERT model, and took the 1,000 least hateful.

For the cross-domain experiment, we used the following five datasets, also used in (Wiegand et al., 2019), plus the dataset of (Qian et al., 2019). The datasets were cleaned in the same manner as RM. The five datasets are imbalanced to different degrees. To control for the effect of dataset imbalance on the results of the cross-domain test, we down-sampled the negative class to match the positive class. Due to computational limitations, we also down-sampled the positive class in the larger sets.

The Kaggle dataset (Kaggle, 2014) contains 312,737 Wikipedia comments, 22,468 of them offensive, labeled with five hate-speech labels (e.g. toxic, abusive, etc). We treat a comment as hate speech (the positive class) if at least one of the five labels is true. We randomly sampled 5,000 comments from each class to form our Kaggle dataset.

The Founta dataset (Founta et al., 2018b) contains 99,799 tweets, 27,037 labeled as abusive, 4,948 as hateful, 14,024 as spam and the rest (53,790) as normal. We sampled 5,000 offensive comments (labeled either abusive or hateful) and 5,000 benign ones. The data itself is available at (Founta et al., 2018a).

The Razavi dataset (Razavi et al., 2010b) contains 1,525 messages, 1,038 non-offensive and 482 "flame", that is offensive texts. We down-sampled the non-offensive class to match the size of the offensive class, giving a total of 964 comments. The data itself is available at (Razavi et al., 2010a).

The Waseem dataset (Waseem and Hovy, 2016a) contains 16,907 tweets, 1,970 labelled with racism, 3,379 with sexism and all the rest (11,559) non-offensive. The online data (Waseem and Hovy, 2016b) contains only tweet ids and labels. We used Twitter's API to recover the text of 795 offensive tweets (sexism and racism) and 3,699 non-offensive tweets. We then down-sampled the non-offensive class to match the size of the offensive class, giving us a total of 1590 tweets.

The Kumar dataset (Kumar et al., 2018) consists of 15,000 Facebook posts and comments, out of them 3,419 tagged as overtly aggressive, 5,296 as covertly aggressive, and 6,285 as non-aggressive. We randomly sampled 5,000 aggressive (overtly and covertly), and 5,000 non-aggressive comments. The authors communicated the data privately after filling out an online application form.

The Offensive Reddit dataset (Qian et al., 2019) consists of 5,020 conversations in which offensive comments are tagged. We sampled 3,230 offensive comments. For the negative class, we sampled 3,230 comments from the political classification task (Washam, 2019) and comments that we harvested from subreddits about fitness and food.

5 Evaluation

We evaluated the quality of our pipeline by augmenting the six hate/abusive speech datasets described in Section 4. The code and datasets can

Train / Test	Kaggle	Founta	Razavi	Waseem	Kumar	Offen. Reddit	CD Avg
Kaggle	0.83	0.75	0.73	0.59	0.59	0.68	0.66
Kaggle + RM	0.85	0.77	0.68	0.58	0.55	0.7	**0.66**
Founta	0.73	0.85	0.56	0.53	0.42	0.67	0.58
Founta + RM	0.76	0.86	0.58	0.55	0.46	0.65	**0.6**
Razavi	0.69	0.62	0.6	0.51	0.45	0.57	0.57
Razavi + RM	0.72	0.72	0.64	0.52	0.48	0.59	**0.61**
Waseem	0.55	0.54	0.54	0.74	0.55	0.61	0.55
Waseem + RM	0.65	0.69	0.5	0.77	0.54	0.59	**0.6**
Kumar	0.61	0.62	0.56	0.64	0.61	0.62	0.61
Kumar + RM	0.76	0.74	0.58	0.63	0.65	0.63	**0.67**
Offensive Reddit	0.67	0.66	0.63	0.58	0.57	0.83	0.63
Offensive Reddit + RM	0.8	0.77	0.65	0.64	0.53	0.78	**0.68**

Table 1: Experiment 1: Cross-domain (CD) and In-Domain (ID) macro-F1 score for the BERT cased-uncased fine-tuned with train dataset (row) and tested on test dataset (column).

Train / Test	Kaggle	Founta	Razavi	Waseem	Kumar	Offen. Reddit	CD Avg
HB+Kaggle	0.91	0.82	0.77	0.74	0.63	0.77	0.75
HB+Kaggle+RM	0.98	0.79	0.74	0.72	0.6	0.74	**0.72**
HB+Founta	0.79	0.92	0.55	0.57	0.44	0.72	0.62
HB+Founta+RM	0.81	0.96	0.6	0.7	0.5	0.71	**0.66**
HB+Razavi	0.73	0.67	0.73	0.62	0.6	0.68	0.66
HB+Razavi+RM	0.83	0.73	0.86	0.68	0.55	0.69	**0.7**
HB+Waseem	0.61	0.61	0.55	0.85	0.59	0.68	0.61
HB+Waseem+RM	0.72	0.73	0.49	0.93	0.54	0.67	**0.63**
HB+Kumar	0.73	0.76	0.71	0.72	0.71	0.73	0.73
HB+Kumar+RM	0.82	0.75	0.66	0.73	0.85	0.68	**0.73**
HB+Offen. Reddit	0.68	0.77	0.7	0.64	0.61	0.92	0.68
HB+Offen. Reddit+RM	0.85	0.81	0.67	0.74	0.57	0.94	**0.72**

Table 2: Experiment 2: Cross-domain (CD) and In-Domain (ID) macro-F1 score for HateBERT (Caselli et al., 2020) fine-tuned with a train dataset (row) and tested on test dataset (column).

Table / Setting	Avg CD	Avg CD+RM	Avg ID	Avg ID + RM
Table 1	0.60 $\pm$0.037	0.64 $\pm$0.034 (**+6.6%**)	0.74 $\pm$0.100	0.76 $\pm$0.085 (**+2.7%**)
Table 2	0.67 $\pm$0.051	0.69 $\pm$0.036 (**+2.9%**)	0.84$\pm$0.088	0.92 $\pm$0.049 (**+9.5%**)

Table 3: Summary of Tables 1 and 2, by averaging over the last column in the Cross-Domain (CD) setting, and over the diagonal in the In-Domain (ID) setting. Improvement in % when using RM is in parenthesis.

Table / Setting	Avg CD	Avg CD+RM	Avg ID	Avg ID + RM
Table 1 30/70	0.63 $\pm$ 0.031	0.68 $\pm$ 0.019 (**+7.9%**)	0.75 $\pm$ 0.079	0.79 $\pm$ 0.049 (**+5.8%**)
Table 2 30/70	0.71 $\pm$ 0.052	0.74 $\pm$ 0.031 (**+5.4%**)	0.82 $\pm$ 0.13	0.825 $\pm$ 0.087 (**+0.8%**)

Table 4: Summary of Experiments 3 and 4 (imbalanced dataset case)

be found at the project's GitHub page (Ilan and Vilenchik, 2022). We ran four cross-domain prediction experiments. In Experiment 1, we repeated the following for every pair of different datasets D_i, D_j. We fine-tuned the BERT-base-uncased model (Devlin et al., 2018) on D_i and tested the

resulting model on D_j. We then repeated the same procedure but now training on D_i+RM, D_i augmented with the RM dataset, and testing on D_j. The results of this cross-domain evaluation are described in Table 1.

Due to limited computational resources (we had one G-Force RTX 3090 GPU), we could not train BERT on the entire dataset. Therefore we broke the large datasets into four random parts (each includes about 1500-2000 comments). The figures appearing in the table are the average over these four-folds. In each fold, we used 80% of the data for training and 20% for validation. In the two small datasets, Waseem and Razavi, we used the entire dataset, repeating four times train (80%) and validation (20%), each time on a randomly sampled 80% of the dataset.

Experiment 2 is identical to Experiment 1, but this time the starting point is the basic pre-trained HateBERT model from (Caselli et al., 2020) fine-tuned with each of the datasets. The results of Experiment 2 are in Table 2.

We also tested the in-domain prediction task in both experiments using 4-fold cross-validation. The results are the diagonal of Tables 1 and 2.

Table 3 summarizes the results of the two tables and shows that augmenting the dataset using RM yielded an overall improvement ranging between 2.7% to 9.5%, depending on the setting (in-domain or cross-domain) and on the initial BERT model. This should be compared to the 5% in-domain average improvement in (Cao and Lee, 2020). Table 4 is the same for imbalanced dataset, and shows higher scores and improvements, especially for Cross Domain task (7.9% for Bert and 5.4% for HateBert).

Experiments 3 and 4 are identical to 1 and 2, respectively. The only difference is that now all datasets D_i and D_j were imbalanced to a 30-70 ratio (30% hate speech) to facilitate a more realistic scenario where the positive class is in the minority. For lack of space, we only give the summary of the results, Table 4. Compared to the balanced setting, we notice that the improvement in the imbalanced setting when using RM is larger in most cases.

All differences in Tables 1,2 between F1 scores after and before augmentation, except one case (in Experiment 4), were verified using a paired t-test and came out significant.

6 Discussion

In this work we have shown that invited abusive speech, which is written humorously, is useful for data augmentation. Our work suggests that humans can produce actual hate speech even without the appropriate psychological conditioning of the brain (such as anger, hate, antagonism, etc). This may hint at some universal properties of hate speech that do not depend entirely on certain emotional states of the mind. We leave this last thought as a gate to further multidisciplinary psycholinguistic research, which may shed more light on the phenomenon of hate speech and how to identify it better using automatic tools.

In trying to get a deeper insight into how exactly did RM help in the cross-domain test, we identified two meta-classes of datasets: randomly sampled datasets with boosting of abusive comments (for example Kaggle and Founta and) and datasets that were selected by topics or key words that were assumed to be assoicated with hate and offensive speech (Waseem and Kumar).

We found that the first group was characterized more by a direct and clear offensive style, while the second group was by a more indirect and fuzzier offensive style. The highest rate of improvement due to RM augmentation was when we trained on a dataset from the second group and tested on a dataset from the first group (improving both in false positive rate and false negative). For example, training on Kumar and testing on Kaggle, we observed an improvement of 25% following RM augmentation. We surmise that the improvement is because RM contains, by its nature, clear and direct offensive comments that complement that missing part in the original dataset. When trained and tested on datasets from the first group, RM mainly contributed to reducing FN-rate, perhaps because it "bridged" the gap between two distributions with its rich and diverse content.

Finally, another angle that our work did not attend to is that of unintended biases (Dixon et al., 2018). If one agrees that unintended biases impair generalizability, then our cross-domain improvement results put forth the premise that augmenting with RoastMe decreased such biases. This point deserves separate in-depth exploration.

Limitations

In our cross-domain evaluation, we did not have sufficient compute power to train the classifier on

the entire dataset. To this end, we broke the date into chunks, training and testing on random chunks each time. This is similar to a cross-validation procedure, which is not necessary in a cross-domain experiment. It may be that the results will change slightly when using the entire dataset.

We did not check the usefulness of other invited hate speech platforms. There is `/r/toastme/`, `r/Rateme/` and probably other platforms where abusive speech is the norm. Therefore we can't say if, in general, such invited hate speech is useful or if we simply got lucky with RoastMe. We surmise that the latter is not the case.

Finally, we ensured that all the datasets in our experiment were balanced (or imbalanced, but to the same degree). We did not check the usefulness of data augmentation using RM for differently imbalanced train and test datasets.

References

Kimberley R Allison, Kay Bussey, and Naomi Sweller. 2019. 'i'm going to hell for laughing at this' norms, humour, and the neutralisation of aggression in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25.

Rui Cao and Roy Ka-Wei Lee. 2020. HateGAN: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, pages 10–30.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

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*, pages 67–73.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018a. Hate and abusive speech on twitter. https://github.com/ENCASEH2020/hatespeech-twitter.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 875–878. IEEE.

Tal Ilan and Dan Vilenchik. 2022. Harald. https://github.com/Talilan/HARALD.

Kaggle. 2014. Toxic comment classification challenge. https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge.

Anna Kasunic and Geoff Kaufman. 2018. " at least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/roastme. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

Kristina Machova, Ivan Srba, Martin Sarnovskỳ, Ján Paralič, Viera Maslej Kresnakova, Andrea Hrckova, Michal Kompan, Marian Simko, Radoslav Blaho, Daniela Chuda, et al. 2020. Addressing false information and abusive language in digital space using intelligent approaches. In *World Symposium on Digital Intelligence for Systems and Machines*, pages 3–32. Springer.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

JT Nockleby. 2000. 'hate speech in encyclopedia of the american constitution. *Electronic Journal of Academic and Special Librarianship*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *EMNLP*.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010a. Flame dictionary. https://www.site.uottawa.ca/~diana/resources/.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010b. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283.

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.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Ravsimar Sodhi, Kartikey Pant, and Radhika Mamidi. 2021. Jibes & delights: A dataset of targeted insults and compliments to tackle online abuse. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 132–139, Online. Association for Computational Linguistics.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *ICML*.

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. 2017. A bayesian data augmentation approach for learning deep models. *Advances in neural information processing systems*, 30.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*, pages 672–680.

Zeerak Waseem and Dirk Hovy. 2016a. 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.

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

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

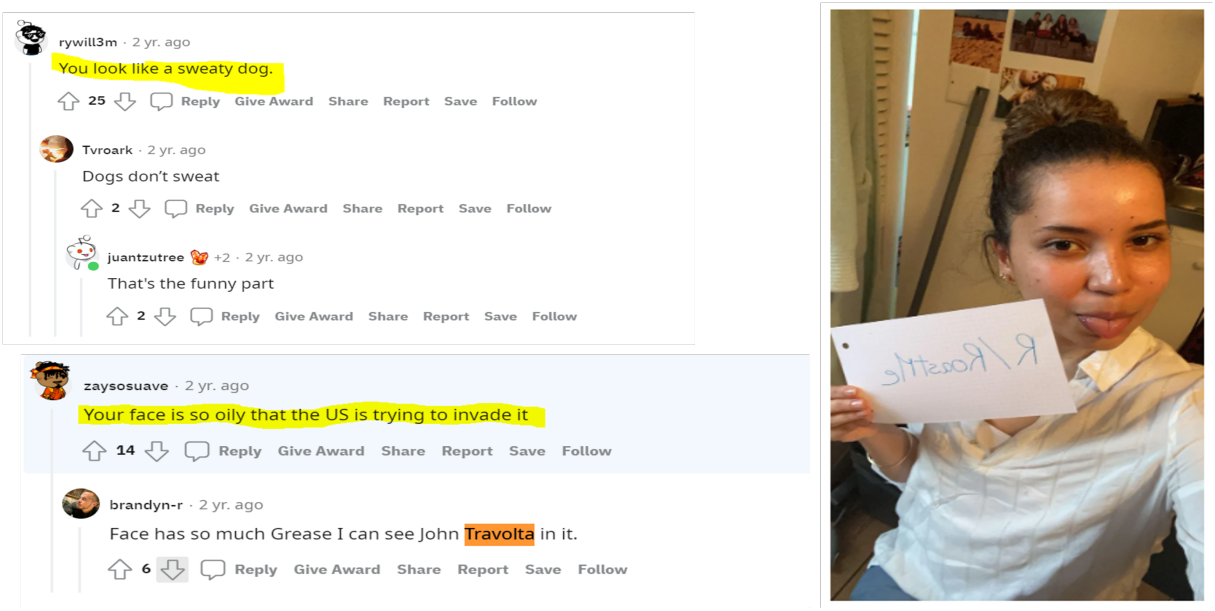

Figure 1: An example for two roasts (highlighter in yellow) from our RoastMe dataset and the photo they were directed to. We Can see that the first comment is more overtly offensive and the second is more covertly offensive. This illustrates the diversity of the roasts, which may have been the key to the improvement of the classification model.