HamiltonDinggg at SemEval-2021 Task 5: Investigating Toxic Span Detection using RoBERTa Pre-training

Huiyang Ding
School of Information
University of Michigan
Ann Arbor, MI 48105
huiyangd@umich.edu

David Jurgens
School of Information
University of Michigan
Ann Arbor, MI 48105
jurgens@umich.edu

Abstract

This paper presents our system submission to task 5: Toxic Spans Detection of the SemEval-2021 competition. The competition aims at detecting the spans that make a toxic span toxic. In this paper, we demonstrate our system for detecting toxic spans, which includes expanding the toxic training set with Local Interpretable Model-Agnostic Explanations (LIME), fine-tuning RoBERTa model for detection, and error analysis. We found that feeding the model with an expanded training set using Reddit comments of polarized-toxicity and labeling with LIME on top of logistic regression classification could help RoBERTa more accurately learn to recognize toxic spans. We achieved a span-level F1 score of 0.6715 on the testing phase. Our quantitative and qualitative results show that the predictions from our system could be a good supplement to the gold training set's annotations.

1 Introduction

Toxic messages remain a small but persistent part of online communications (Fortuna and Nunes, 2018; Jurgens et al., 2019). NLP methods have been developed to identify these comments, often relying on deep-language models (Vidgen et al., 2019). However, the part of the message that is specifically toxic is often unknown. Such information is useful not only for validating and explaining the judgments of models (Carton et al., 2018), but can also be useful for moderators to use when making decisions and working with these models in their deployment (Carton et al., 2020; Liu et al., 2021). This paper describes our model[1] and error analysis for SemEval-2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021).

Our model uses a deep learning approach to identify which tokens are toxic. The approach is motivated by two strands of prior work showing (1) that large language models can effectively serve as sequence-to-sequence (seq2seq) models and (2) that pre-training on a similar task can improve downstream performance (Phang et al., 2018; Gururangan et al., 2020). Here, we treat the toxic-span detection tasks as a seq2seq task, where given a sequence of tokens, the model outputs per-token judgments of whether the token is in the toxic span. Given the limited training data for Task 5, we increase our training data by generating a silver-standard set of span judgments from LIME explanations (Ribeiro et al., 2016) fropm a model trained to recognize toxic and non-toxic language. These additional judgments are intended to help the model learn the basic span recognition task and identify general toxic language, before fine-tuning on the Task 5 data.

2 System Description

Our core system relies on a standard RoBERTa model (Liu et al., 2019) that is trained on a sequence-to-sequence task in two phases. The first phase pretrains the model with heuristically-created spans, gathered from Reddit comments labeled for their toxicity. The second phase fine-tunes this model on the organizer-provided data. Figure 1 shows the overview of the system. All the data used in the paper is in the English language.

2.1 Pretraining to Recognize Toxicity

To identify toxic spans, we hypothesize that pre-training the RoBERTa model on a similar task would lead to better downstream performance. Therefore we generate a similar dataset (silver dataset) to the training data (gold dataset) and heuristically label it with spans by using LIME (Ribeiro et al., 2016) on a toxicity classification task.

[1]The code is available at https://github.com/davidjurgens/offensive-span-detection.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 263–269
Bangkok, Thailand (online), August 5–6, 2021.

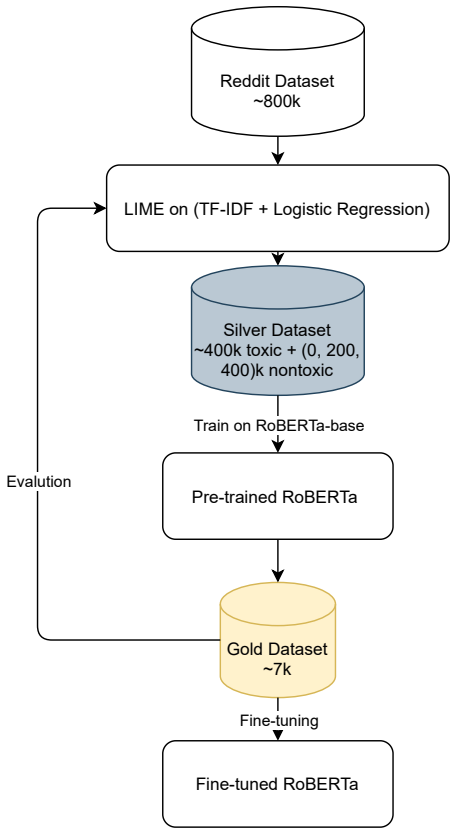

Figure 1: Diagram of our data and architecture. The central hypothesis tested is whether pre-training a RoBERTa model on machine-generated rationales for toxicity could improve performance.

Data Silver data was drawn from a sample of all Reddit comments made between January to June 2018. As social media data, these comments contain similar lexical and syntactic patterns as the social media comments data as the gold standard, which was made on the Civil Comments platform. Prior work has shown that pre-training RoBERTa models to recognize this type of social media data improve downstream performance (Nguyen et al., 2020). However, Reddit posts can vary substantially in their length. To avoid introducing confounding effects from pre-training a model on posts of substantially different lengths, we compute the Inter-quartile Range (IQR) of the lengths of Reddit comments and remove all comments identified as outliers. This process effectively removes very long or very short comments. Ultimately, the mean number of words in the training data and Reddit data are roughly similar: 35.87±34.92 words (mean and standard deviation) in the training data, compared with 36.79±30.57 in the Reddit data.

Identifying Toxic Comments The Reddit data contains a mix of toxic and non-toxic conversations, which we aim to use for training. To identify toxic conversations, we use the Perspective API to label all comments in the dataset (Wulczyn et al., 2017). The API returns a continuous score reflecting the degree of a comment's toxicity. This toxicity score is then converted into a binary label to use in training a LIME model to generate rationales for why a comment is (or is not) toxic. We follow the insights from Hua et al. (2020) and set a threshold of 0.7, above which a comment is considered toxic and 0.3, below which the comment is non-toxic. These thresholds were intended to help create easy examples of toxic language for generating rationales as a way of scaffolding the learning for the downstream task. This process led to a labeled dataset of 288.5M comments with binary toxicity labels, of which 9.4% were labeled toxic.

Generating Heuristically-Labeled Toxic Spans Our final silver dataset is created by sampling comments from the larger labeled Reddit comments and using a LIME model to generate toxic span labels. LIME is a form of interpretable machine learning that explains the decisions of a classifier using a local approximation to identify which features led to a classification decision. Here, we use a simple logistic regression (LR) model trained on TF-IDF features and use LIME to generate a rationale of the classifier's decision which identifies which words are contributing to the toxicity decision. The underlying LR model is trained on a balanced sample of 800K toxic/non-toxic comments (not the silver data). This balanced sample is derived in the same way as silver data. The model's hyperparameters were tuned using 10-fold cross-validation, with thae learning rate of 0.01 and strength of the regularization (C) at 1 under L1 loss. In a test of a held-out 200K instances, the model attained a binary F1 of 0.985.

Our silver data is created by generating LIME explanations using the trained LR classifier on a separate 800K comments balanced between toxic and non-toxic. This size is roughly 100x the Task-provided data. In generating explanations, LIME assigns local weights to each token on its weight to drive the correct prediction. To create toxic spans from these continuous-valued weights, we apply a threshold above which we consider the token as the toxic span. The threshold was identified by generating LIME explanations for all of 8629

documents from the Task's training data and then choosing the threshold that maximized the Span F1 between the 8629 training documents' toxic spans and the discretized LIME explanations, using a grid search with a step of 0.001 in [0.05,0.50]; the final threshold was set to 0.169.

2.2 Model Training

Our model uses a common RoBERTa base and differs according to which data the model is trained on. The pre-training setup trains a RoBERTa model on a seq2seq task where the input sequence of tokens generates a binary sequence denoting whether the input token was inside or outside of a toxic span. Due to the dataset sizes, pre-training was done for one epoch. The fine-tuning setup starts from either the off-the-shelf RoBERTa parameters or from a RoBERTa model initialized through silver-data pre-training. This model is trained in the same way as in pre-training on a binary seq2eq task using the Task-provided data. Models are fine-tuned for 10 epochs and parameters are chosen using the epoch with the best performance on the trial data.

In internal testing, we compared models that have been pre-trained, fine-tuned, or both, using varying amounts of silver data. All hyperparameter choices are reported in the Appendix A.

3 Results

Our best model attained a Span F1 of 0.672, and although close in score to the top result (0.708), was ranked 30 in the Task. Surprisingly, this best-performing model did *not* make use of the pre-training on silver data. To better understand the performance, we ran two follow-up analyses to test how different strategies for training affected performances and an error analysis for what were common themes in errors.

3.1 Does Pre-training Make a Difference?

In assessing the impact of pre-training, we analyze the submitted model along with five other models: (1) a fine-tuned model using a batch size of 8, (2) a pre-trained only model that makes no use of the Task data, and (3-5) pre-trained and fine-tuned models that use different amounts of silver data. The performances of all models are shown in Table 1.

For the initial comparison, we contrast the fine-tuned model (Table 1, Row 1; denoted FT) with the pre-trained and fine-tuned model on all silver data (Table 1, Row 6; denoted PTFT). Both models agree on 1281 (65%) of the 2000 test instances. For these agreed cases, both models attain a Span F1 of 0.776—higher than either models regular performance. In these matching predictions, the ground-truth spans have an mean length of 1.13 tokens, mainly concentrated on commonly-labeled offensive words, like "morons", suggesting that both models are adept at identifying overtly toxic words. In contrast, for the 21 test documents whose spans have $\geq$5 words, both models perform poorly with a Span F1 of 0.3489 for the FT model and 0.2077 for the PTFT model.

Model	Batch Size	Silver Data	F1
FT	8	N/A	**0.675**
FT	16	N/A	0.672†
PT	8	400k/400k (1:1)	0.613
PT + FT	8	400k/0 (1:0)	0.660
PT + FT	8	400k/200k (2:1)	0.660
PT + FT	8	400k/400k (1:1)	0.659

Table 1: Performance at recognizing toxic spans (Span F1) for models trained on just the Task-provided training data (baseline), Pre-Trained (PT) on different amounts and ratios of silver data, and Fine-Tuned (FT) on training data. Ratios denote the number of non-toxic:toxic examples. † is the model submitted to the Task.

The FT model performs considerately better than PTFT model on documents with $\geq$5 tokens labeled as ground-truth spans, which is likely due to differences between the LIME-labeled data and the Task's training data. For example, the LIME model generates spans $\geq$5 tokens in only 693 of the 800K silver context, suggesting LIME tends to give shorter toxic span labels.

This bias affects the downstream model performance in the test set where 29 of the 2000 test contexts have a span of $\geq$2 consecutive toxic words. In those contexts, the FT model achieves a mean Span F1 of 0.523 while the PTFT model has only 0.369. Indeed, the FT model produces spans (average span length: 19.0345) that are $\sim$224% longer than the PTFT model (average span length: 8.4828). This difference is more obvious than the overall prediction results, as seen in Table 2.

In the remaining cases where the FT and PTFT model predictions differ, the FT model has 296 predictions with a better Span F1 score, of which 216 predictions have longer span. For example, in test context 100, "Stupid is as stupid does Gump was right," both of the "stupid" tokens are highlighted

Span Size	FT model	PTFT model
Characters	9.30 ± 6.44	7.05 ± 3.48
Tokens	2.14 ± 1.72	1.61 ± 1.00

Table 2: Differences in model prediction length shows that pretraining on LIME-generated toxicity rationales (PTFT) generally produces shorter spans at both span and token levels.

by the FT model, while the PTFT model only labels out the first "stupid".

However, the tendency of the FT model to predict longer spans does not always yield higher performance. The PTFT model has 237 predictions with better Span F1 scores.

The FT model has a longer span prediction in the 202 of the 237 predictions, but the qualitative results are very different from the above-mentioned example. In multiple cases when the ground-truth spans are empty, the PTFT model can also predict empty spans. However, the FT model has a lower tendency to predict empty spans in those cases. For example, in non-toxic test context 3, "The parallels between the ANC and the Sicilian Mafia are glaring...", the FT model labels "Sicilian" as toxic, while the PTFT model output is (correctly) empty.

Looking at contexts where there are no underlying toxic spans, the two models perform slightly differently. There are 394 out of 2000 test contexts with empty ground-truth spans. For those contexts, the FT model only gets a mean Span F1 score of 0.058, while the PTFT model gets 0.079.

In contrast to the FT model, the PTFT model has less-accurate predictions on the overtly/commonly toxic spans. For example, there are 430 total "stupid" or "stupidity" related words labeled as toxic by the ground-truth spans. The FT model is able to label 383/430 as toxic, while the PTFT model only labels 331/430. As we know, words like "stupid" can be more contextually-sensitive when compared to other common offensive words. They could be used in a toxicity-neutral way in many contexts. In the PTFT model's pre-training phase, we fed 400,000 non-toxic documents for the RoBERTa model. These non-toxic documents supplied more non-offensive context for certain toxic words than the small-sized gold dataset. The extra contextual information learned by the pre-trained model can somehow decrease the performance of the PTFT model.

3.2 Common Themes in Errors

From the error analysis in the above section, we have noticed that the PTFT model does not perform well when it comes to predicting long toxic spans, empty toxic spans, and toxic phrases. With a deeper dive into the differences between PTFT model predictions and ground-truth spans, we can get a better sense about how is our PTFT model doing and what insights could be recommended for annotators.

In order to perform both qualitative and quantitative error analyses, we randomly sampled 200 test contexts where predictions from the PTFT model do not conform with ground-truth labels. The overall mean Span F1 score on those mispredictions is 0.289; compared to the random sample with a mean of 0.274. We categorize the mistake types from the sampled contexts and provide examples in Table 3.

Model Errors The first four categories (Category 1–4) in Table 3 demonstrate the mistakes made by the PTFT model and ground-truth annotations.

Category 1 shows where the PTFT model identifies valid toxic spans not present in ground truth, which accounts for 101 (50.5%) of the model errors in the 200 sampled contexts. In these cases, annotators marked nothing as toxic in 58 contexts. However, most of the overlooked toxic spans are overly-common toxic words like Examples 491 and 1374 in Table 3.

In comparison, in Category 2, there are 81 out of 200 sampled contexts with unmarked toxic span labels in the PTFT model output, in which the PTFT model produced an empty span or an incorrect span as toxic in four cases. Example 642 in Table 3 shows a typical case were a relative rare toxic word, "caca," is overlooked by the model. In the remaining cases, the PTFT model has shorter predictions than the ground truth toxic spans, matching the low performance on predicting longer toxic spans (§3.1). In some cases, when multiple toxic spans exist in the same document, ground truth only labels one or two spans of them (e.g. Examples 1852 and 1486). While in other cases, ground truth would label more toxic spans (e.g. Example 346). The inconsistencies from ground truth annotations occur more when it comes to the multiple-word labeling, which will be described more in the next analysis for Category 5.

The missing labels from both scenarios are

Category	%	Examples
1. Toxic not labeled by ground-truth spans	50.5	491. this guy is an idiot . I'll disclose I'm male.... 1374. That is a bunch of horse sh*t . 1776. **Stupid** leftists. Stupid Democrats. I sincerely..
2. Toxic not labeled by the PTFT model	40.5	1852. That's because Floater **sucks** . That dudes voice is so **fucking** annoying! 1486. uhhh Hillary Clinton is a serial killer and **thief** 346. ...Drivers in Anchorage are lazy, **stupid** and **impatient**... 642. What a load of **caca**!! ...
3. Not toxic but labeled by ground-truth spans	0.5	876. Ergo, America is overrun with the **insane**.
4. Not toxic but labeled by the PTFT model	13.5	273. Extremist Muslims want to kill infidels. Moderate Muslims cook the extremists a hot lunch. 1802. either you can afford to lose $1300. in an hour or you're a moron
5. Inconsistent multiple-word/phrase annotation by ground-truth	4	773. Very true. Still sick **bastards** . 1496. Trump is an **impulsive** **idiot** . He will get us all killed. 1776. Stupid leftists. **Stupid** Democrats. I sincerely believe... 1447. Brooks, would you please join the **damn** **Democrat** party and be done with it ?
6. Inconsistent word annotation by ground-truth	41.5	968. ok then you dont use gasoline, plastic or such anything else right??? ya **hypocrite** bs **stupidity** as usual 348. ...Hawaii Democrats deny ordinary citizens their constitutional right to self-defense with firearms, including concealed carry... Hypocrites !
7. Inconsistent repeated word annotation by ground-truth	2	413. There is a difference between being tolerant and being **stupid** . She and her supporters want America to be **stupid** . 137. I cannot fathom the **stupidity** of McCain believing any cooperation from the ... The man must be the stupidest person in D.C.
8. Inconsistent sentence annotation by ground-truth	12.5	894.**You are a** **nut ball.** 1374.That is a bunch of horse sh*t .

Table 3: An error analysis of 200 predictions of our PTFT model relative to the ground-truth span. All of the sample contexts in the Examples column begin with the context index starting from 0. The first four categories show the differences in between the toxic spans labeling from PTFT model (shaded in blue) and the ground-truth span (bolded). The last four categories show the commonly-seen inconsistencies existing in the ground-truth annotations. Within the last four categories, examples are shown in pairs. Within each pair of examples, the upper example shows the suggested toxic spans based on the majority decisions made by the test context annotation (shaded in yellow) and the ground truth labeling decision by the annotators (bolded). In contrast, the lower example within each category shows where the consistency of annotation breaks compared to the upper example.

straight-forward in their formats and contexts. Besides, there are 20 contexts that both the ground-truth spans and PTFT model missed the toxic span partially or completely.

Generally, ground truth annotation seldom labels non-toxic spans as toxic (Category 3). On the contrary, it is common for our PTFT models to make mistakes on labeling non-toxic spans (Category

4). It happens mostly in the cases when the PTFT model misinterprets the context (e.g. Examples 273 and 1802).

Inconsistencies in Ground Truth Labels The last four categories (5–8) in Table 3 show common inconsistencies in annotation decisions, which we hope could aid in improving consistency in future work.

In Category 5, the standard for spans labeling is not consistent for which words are included in the toxic phrase. In some cases, when a sentence is fairly short (< 5 words) and contains toxic words, the ground truth annotation would label out the adjective used for describing the trailing noun (e.g. Example 1469). However, in other cases (e.g. Example 773), the standard would change by skipping the adjectives. Moreover, this inconsistent annotation also occurs in the case when the underlying nouns are almost the same (e.g. Examples 1447 and 1776).

Categories 6 and 7 comprise the majority of the inconsistencies in the annotation standards by the ground truth. These inconsistencies commonly manifest for frequent toxic words. For instance, in Examples 968 and 348, both of the "hypocrite(s)" should be toxic given the context and there is no more than one other toxic word within the document. The omission of common toxic words is the major source for this category. In many cases, the subtle variations of the document context would make it even harder to maintain a unified standard across different annotators. Hence, the introduction of some model-based labeling (or checking) could greatly improve the inconsistencies of this case.

In the sampled contexts, 25 documents consist of only one sentence. Annotators varied in how much of these contexts to label (Category 8), occasionally marking the entire sentence as offensive (20% of these single-sentence contexts), as in Example 894. However, in a few cases (4 of 25), annotators labeled nothing as toxic (e.g. Example 1374). Interestingly, 9 of 25 cases where ground truth either labels the entire sentence or nothing, our PTFT model is able to identify the toxic word(s), suggesting the model is still effective for short contexts.

4 Discussion and Future Work

Based on error analysis, our PTFT model suffers from low performance when generating predictions on non-toxic contexts or long toxic spans. A modified error function that rewards for edge-case scenarios can potentially improve the PTFT performance. Moreover, during the pre-training, we applied a simple-cutoff on the local weights to make labeling decisions for LIME explanations. The cutoff was determined solely based on evaluations with the Task data. If the LIME labeling could introduce more robust variants in the loss evaluation, the silver data span labeling might be more representative of the Task data's annotation logic.

Through comparing silver data with the gold data, we find the toxicity of some words is influenced by the broader linguistic environment. While the silver and gold data both consist of online comments, their time spans and topics are very different. The gold data uses contexts from 2015-2017 and has a concentration on political news; while the silver data covers 6 months of 2018 with no focused topics. Our qualitative analysis finds that the addition of non-toxic examples in the silver data influenced the model to consider overly common toxic words less toxic than they were in the gold data. Future work is needed to identify the optimal ratio of the toxic and non-toxic samples and to address domain/register differences in the data.

Last, the current approach could invite several natural improvements. For example, in the pre-training phase, we used TF-IDF embedding and logistic regression for the base of LIME explanations. This combination was chosen for its efficiency in the LIME training phase. However, many other embedding and model combinations rendered much better classification results, which may generate better rationales for pre-training.

5 Conclusion

We presented our system for SemEval-2021 Task 5 on Toxic Span Prediction. Our initial approach used explainable machine learning (LIME) to generate a heuristically labeled span dataset, which was used to pre-train a RoBERTa model to recognize toxic spans. However, our results show that when fine-tuned on the task data, the resulting model generates slightly shorter explanations and ultimately performs slightly worse than a model trained only on the Task's training data—likely due to bias towards shorter spans generated by LIME. In our subsequent error analysis, we show that the majority of our model's errors (50.5%) are associated with missed annotations in the ground truth, suggesting that actual model performance may be higher in practice.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No 185022.

References

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 95–106.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020. Towards measuring adversarial twitter interactions against candidates in the US midterm elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 272–282.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *arXiv preprint arXiv:2101.05303*.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Hyperparameter	Value
learning rate	5e-05
number of epochs	1
weight decay	0
model	RoBERTa-base
dropping probability of the dropout layer	0.5
learning rate scheduling	linear
optimizer	AdamW
warmup steps	300
random seed	42

Table 4: Hyperparameters of pre-trianed model

Hyperparameter	Value
learning rate	5e-5
number of epochs	10
weight decay	1e-2
model	RoBERTa-base
dropping probability of the dropout layer	0.5
learning rate scheduling	linear
optimizer	AdamW
warmup steps	0
random seed	42

Table 5: Hyperparameters of fine-tuned model

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the Web Conference*.

A Hyperparameter Settings

Tables 4 and 5 respectively show the hyperparameter settings for the fine-tuned (FT) model and the model pre-trained on silver data and then fine-tuned on the training data (PTFT).