Probing the impact from fine-tuning on BERT's ability in understanding toxic speech detection

Lawrence Lai
University of California, Berkeley
lawrence.c.lai@berkeley.edu

Abstract

This paper investigates the ability of the contextual representations learned and generated by BERT, a Transformer-based pre-trained language model, in understanding toxic speech detection before and after fine-tuning. More concretely, we aim to shed light on the following question: How does the linguistic knowledge captured and encoded by the contextual embeddings during pre-training and fine-tuning impact the performance of the downstream task of toxic speech detection? We conduct our experiments using probing tasks, which are supervised classification tasks designed to analyze the linguistic knowledge captured and encoded in contextual embeddings, based on a simple linear probing model architecture and task datasets on topics often associated with toxic speech. We find that pre-trained contextual embeddings already have the capacity to identify on- and off-topic mentions in the presence and absence of toxicity. Our analysis reveals that the fine-tuned contextual embeddings do not have the ability to distinguish individual toxic components of a post, which suggests that toxicity of individual words is not encoded in the contextual embeddings even after fine-tuning. We also observe that fine-tuning yields varying performance impact on contextual embeddings across all hidden layers where some of the best performing contextual embeddings are found in earlier hidden layers deep in the Transformer stack of a BERT model. *Warning: This paper contains samples of texts that are offensive and/or hateful in nature.*

1 Introduction

In the ever-expanding digital landscape, the proliferation of online communication platforms has facilitated unprecedented levels of connectivity, enabling individuals from diverse backgrounds to engage in discourse on a global scale. While this connectivity has undeniably enriched the exchange of ideas and perspectives, it has also given rise to a problem: the pervasive presence of toxic speech within these virtual spaces. Toxic speech, encompassing hate speech, harassment, and various forms of harmful language, poses significant challenges to maintaining a healthy and inclusive online environment.

To counter toxic speech, there has been an active line of work to identify and filter out toxic speech on the Internet. Most published work on toxic speech detection focuses on classifying an entire comment or document as either toxic or non-toxic, leveraging contextual representations generated by deep neural networks, such as the Transformer-based BERT. However, little investigative work has been done to understand the rationales behind such determination. As such, our objective is to understand how information encoded in contextual embeddings, with and without fine-tuning, facilitates the task of toxic speech detection.

Prior works on examining the knowledge captured by contextual embeddings often rely on analyzing attention patterns and/or classifier-based probing tasks. In our work, we focus on using classifier-based probing tasks to generate insights on the impact of fine-tuning on contextual embeddings in performing the following subtasks relevant to toxic speech detection across six topics often associated with toxic speech: (a) identify words belonging to a topic regardless of toxicity, in the presence of toxicity and in the absence of toxicity, and (b) identify words belonging to a topic that are used in a toxic manner across toxic and non-toxic comments.

For the probing experiments, we generate our own probing task datasets and employ a simple linear probing model architecture to extract information from contextual embeddings as generated from all hidden layers from the original BERT model and two fine-tuned BERT models using post-level toxicity only and both post-level toxicity and rationales. To understand the impact from fine-tuning, we compare and contrast the performance of contextual embeddings on all probing tasks and topics and review the performance differentials across different hidden layers from the three sets of contextual embeddings.

2 Background

The rise of contextualized word representations and their state-of-the-art performance on downstream tasks have spurred researchers to understand and compare these representations in a systematic manner, especially

using probing tasks in a variety of forms at the subword-, word- and sentence-levels. Conneau et al. (2018a) propose SentEval, a toolkit used to evaluate universal sentence representations via binary and multi-class classification, natural language inference and sentence similarity. Wang et al. (2018) present GLUE, a suite of nine tasks designed for probing models for understanding specific linguistic phenomena. Conneau et al. (2018b) introduce ten probing tasks to capture simple linguistic features of sentences such as surface, syntactic and semantic information. Tenney et al. (2019) introduce an edge probing framework along with a suite of sub-sentence tasks to investigate a range of syntactic, semantic, local and long-range phenomena.

With respect to toxic speech detection, there is also an active line of work where researchers collect, annotate and publish a wide variety of datasets with different annotation schemes for further scientific explorations. Majority of the published datasets, such as the ones by Zampieri et al. (2019), Founta et al. (2018), Davidson et al. (2018), etc. are annotated with binary (e.g. toxic or non-toxic) or multi-class (e.g. hate speech, offensive, normal) labels. Very few of the published datasets, such as the ones by Pavlopoulos et al. (2021) and Mathew et al. (2021), contain word-level annotations associated with toxicity rationales.

Our work is inspired by the culmination of the previous works on edge probing framework by Tenney et al. (2019), HateXplain dataset by Mathew et al. (2021), Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. (2016) and layer-wise analysis of Transformer representations by van Aken et al. (2019).

3 Methods

In this section, we describe our approach to investigate the impact of fine-tuning on a model's ability in understanding what makes a speech toxic. More concretely, we first create our control and experimental models by fine-tuning them ourselves or sourcing it from an existing implementation. We then examine the ability for the contextual representations to encode information about a token's role in determining the toxicity of a post via probing experiments.

3.1 Dataset for fine-tuning

We use HateXplain as introduced by Mathew et al. (2021), which is a benchmark dataset for explainable hate speech detection. The dataset consists of 19,229 tweets from Twitter and posts from Gab (collectively referred to as posts hereafter), which are split into train, validation and test sets in the ratio of 8:1:1. Each post is anonymized by replacing the usernames with <user> tokens and is tokenized at the word level by removing all punctuations except for apostrophes. Each post is annotated from three different perspectives, including a three-class classification (hate speech, offensive, or normal), the targeted community and the rationales behind each annotator's labeling decision. Rationales are only provided in the event that the post is considered as hate speech or offensive by the majority of the annotators.

For the purpose of this study, we simplify the original three-class classification into a binary classification by re-labeling hate speech and offensive as toxic and normal as non-toxic. Moreover, for each post, we derive a single ground truth label by employing majority voting. For example, if a post is labeled as either hate or offensive by two out of three annotators, then we assign a ground truth label of toxic for the post. With majority voting, each split contains 40.6% toxic posts and 59.4% non-toxic posts. We follow the same approach in deriving a single ground truth label for the rationales. For instance, if a word from a post is labeled as part of the rationales by two out of the three annotators, then we deem this word as a valid part of the rationales. Non-toxic posts do not have any rationales annotated. As such, all words from non-toxic posts are deemed to be non-toxic rationales.

3.2 Baseline and experimental models

To understand the impact of fine-tuning, we train a baseline model ("Model 1") and an experimental model ("Model 2"). We leverage the BERT base model (uncased) for sequence classification as implemented by Huggingface to fine-tune both models using the binary labels. We use the train, validation and test splits as provided in the original dataset without any modifications. In addition, we also use the BERT model as fine-tuned by the authors of HateXplain using both binary labels and rationales ("Model 3") as another experimental model for comparison purposes.

For Model 1, we freeze all layers except for the final sequence classification layer during training, yielding 769 trainable parameters and 109,482,240 non-trainable parameters. We experiment with all optimizers as implemented by Keras and learning rates from 0.00001 to 0.001 using the train and validation sets. The best iteration achieves a weighted average F1 score of 0.63 on the test set after training for 18 epochs using the RMSprop optimizer with a learning rate of 0.001.

For Model 2, we unfreeze all layers during training, yielding 109,483,009 trainable parameters. We

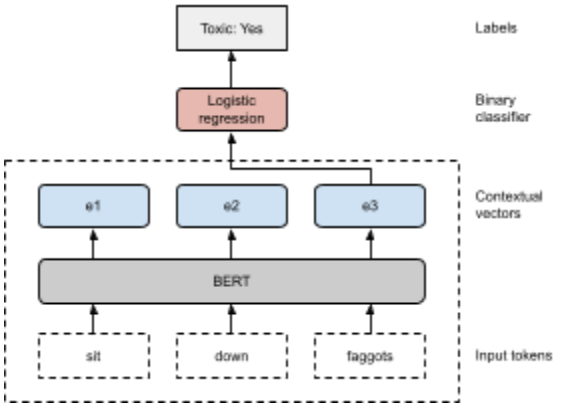

Figure 3.4. The architecture for probing tasks. All parameters inside the dashed line are fixed while we train the logistic regression classifier to extract information from the contextual vectors.

experiment with all optimizers as implemented by Keras and learning rates from 0.00001 to 0.001 using the train and validation sets. The best iteration achieves a weighted average F1 score of 0.77 on the test set after training for 6 epochs using the AdaMax optimizer with a learning rate of 0.0001.

For Model 3, we leverage the implementation as shared by the authors of HateXplain on HuggingFace. The model is a BERT model fine-tuned using post-level toxicity and ground truth attention. Ground truth attention is generated by converting word-level rationales into boolean vectors, taking the average across annotators, and normalizing the resultant vector through a softmax function with a temperature parameter. For non-toxic posts, each element in the ground truth attention is replaced with 1/(sentence length) to represent uniform distribution. We do not perform any additional fine-tuning on this model. To evaluate this model consistently with Models 1 and 2, we remap the three-class predictions (hate speech, offensive and normal) to binary predictions (toxic vs non-toxic). This model achieves a weighted F1 score of 0.78 on the test set.

3.3 Dataset for probing

We generate the probing dataset using the test set from the original dataset. Each post from the test set is broken down into multiple samples, each of which consists of one token (i.e. subword) and one ground truth binary label indicating whether the token's corresponding word is included as part of the rationales for the annotators' labeling decisions.

For posts that are deemed toxic, the corresponding rationales are used to generate the ground truth labels. For posts that are deemed non-toxic, all words within such posts are labeled as non-toxic. For each sample, we then extract the contextual embeddings from all 13 hidden layers (including the first embedding layer) for the individual token using all three models. To extract the contextual embeddings corresponding with post tokens correctly, we create a mapping between post token indices and BERT subword token indices across all samples in our test set. For a post token that is mapped to multiple BERT subword tokens, we generate one contextual embedding by taking a simple average of the subword contextual embeddings.

3.4 Probing model architecture

To restrict the expressive power of our probing models, we opt to use a simple logistic regression model instead of a multilayer perceptron model as proposed in the framework by Tenney et al. (2019). The objective is to focus on what information can be extracted from the contextual embeddings with as few trainable parameters as possible during training. Figure 3.4 provides an illustration of our probing model architecture.

3.5 Topics

For our probing experiments, we investigate the following six topics, as shown in Table 3.5. For each topic, we look for mentions of words that are related to and/or used in a toxic manner towards women as deemed by annotators in the test set. We also do our best to identify as many associated terms as we can based on definitions from various online dictionaries and the context in which they are used.

3.6 Tasks

For each topic, we create multiple probing tasks inspired by the findings from the analyses on model errors and LIME coefficients, as included in the Appendix. These tasks and their respective data generation processes are described below. The more represented class is downsampled to arrive at a balanced task dataset. We denote "toxic token" as a token with a role in making a post toxic and "non-toxic token" as a token without a role in making a post toxic as deemed by annotators from the original dataset. All tokens from non-toxic posts are assumed to be non-toxic tokens. An illustrative example of how a post and its tokens are defined is shown in Table 3.6(a). The number of positive and negative samples broken down by topics and tasks available for our probing datasets is illustrated in Table 3.6(b).

Task A: On-topic tokens vs. off-topic tokens. We generate the positive samples by subsetting the dataset to tokens with on-topic mentions and the negative

samples by subsetting the dataset to tokens without on-topic mentions. This task probes the ability for contextual embeddings to encode information for recognizing the topic in general.

Task B: On-topic tokens in toxic posts vs. off-topic tokens in toxic posts. We generate the task dataset by subsetting the positive and negative samples from Task A to ones from toxic posts. This task probes the ability of contextual embeddings to encode information for recognizing the topic in toxic posts only.

Task C: On-topic tokens in non-toxic posts vs. off-topic tokens in non-toxic posts. We generate the task dataset by subsetting the positive and negative samples from Task A to ones from non-toxic posts. This task probes the ability of contextual embeddings to encode information for recognizing the topic in non-toxic posts only.

Task D: Toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in non-toxic posts. We generate positive samples by subsetting the positive samples from Task A to toxic tokens from toxic posts and negative samples by subsetting the positive samples from Task A to non-toxic tokens from non-toxic posts. This task probes the ability of contextual embeddings to encode information for distinguishing between toxic on-topic tokens in toxic posts and non-toxic on-topic tokens from non-toxic posts.

Task E: Toxic on-topic tokens in toxic posts vs. non-toxic on-topic tokens in toxic posts. We generate positive samples by subsetting the positive samples from Task A to toxic tokens from toxic posts and negative samples by subsetting the positive samples from Task A to non-toxic tokens from toxic posts. This task probes the ability of contextual embeddings to encode information for distinguishing between toxic and non-toxic on-topic tokens in toxic posts only.

3.7 Experiments

For each experiment, we perform the following procedures 10 times: We split the task dataset into train and test sets using scikit-learn's StratifiedGroupKFold method to ensure no data leakage across the two splits, such that each post can only be in either train set or test set. We train a probing classifier ("PC") for each model by using the train set and then evaluate using F1 score on the test set. We then average the F1 scores on the test set as our evaluation metric for comparison purposes. Our benchmark is a PC that always predicts the positive class on a balanced task dataset that yields an average F1 score of 0.67 on the test set.

Topic	Variants of terms
Profanity	Common swear words that do not target specific communities such as "ass", "fuck" and "shit"
Women	"bitch", "cunt", "dame", "divorcee", "female", "feminism", "femme", "girl", "hoe", "lady", "mother", "prostitute", "queen", "sister", "slut", "whore", "wife", "witch", "woman" and "yentas"
LGBTQ	"dyke", "fag", "gay", "homo", "homosexual", "lesbian", "lgbt" and "queer"
African	"african", "black", "colored", "coon", "ghetto", "mooncricket", "nigger", "negro", "sheboon" and "spade"
Jewish	"antisemitism", "chabad", "goy", "hebrew", "heeb", holocaust", "israel", "jew", "kike", "rabbi", "semite", "shekel" "synagogue", "yid" and "zionism"
Islam	"allah", "burqa", "halal", "hamas", "hezbollah", "islam", "jihad", "mecca", "medina", "mohammed", "moslem", "mosque", "muslim", "mussie", "muzrat", "muzzie", "umma" and "zakat"

Table 3.5: A list of topics under investigation and variants of associated terms that are used to generate the task datasets

Post	On-topic post	Toxic post
dorothy a dirty **slut**	Yes	Yes
meryl streep in little **women**	Yes	No
sit down faggots	No	Yes
<user> dm me	No	No

Table 3.6(a): An illustration of how a post and its token are defined with respect to the topic of women. Tokens in bold are considered on-topic tokens. Tokens highlighted in red are considered toxic tokens as deemed by annotators.

Topic	Task A	Task B	Task C	Task D	Task E
Profanity	411:46K	297:27K	114:19K	129:114	129:168
Women	526:46K	325:27K	201:19K	153:201	153:172
LGBTQ	354:46K	230:27K	124:19K	190:124	190:40
African	673:46K	473:27K	200:19K	397:200	397:76
Jewish	368:46K	297:27K	71:19K	206:71	206:91
Islam	355:46K	262:27K	93:19K	185:93	185:77

Table 3.6(b): The number of available positive and negative samples by topics and tasks prior to downsampling to create the probing datasets

4 Results and discussion

In this section, we present a summary of findings from our probing experiments. All figures with two decimals in parentheses are F1 scores unless presented otherwise (e.g. percentages). A detailed discussion of the experiment results is provided in the Appendix.

Finding 1: Contextual embeddings from the original BERT model encode sufficient information needed to distinguish between on- and off-topic mentions. We observe that Model 1, which is trained using the original BERT model without any fine-tuning, consistently outperforms our benchmark across Tasks

A, B and C and all topics, performing the best at profanity mentions (0.99) and the worst at women mentions (0.97). We hypothesize that the slight performance difference may be attributable to the prevalence of each topic mentioned (i.e. high for profanity and low for women) in the dataset on which the original BERT model is pre-trained.

Finding 2: The presence and absence of toxicity have minimal impact on the ability of contextual embeddings in encoding information necessary to discern between on- and off-topic mentions. When we subset our task dataset to toxic and non-toxic posts only in Tasks B and C, respectively, we observed minimal performance differences (-0.00 to +0.01). This pattern holds across all three models with and without fine-tuning. This observation suggests that the presence of toxicity does not impact the ability of contextual embeddings to encode information needed to distinguish between on- and off-topic mentions. We hypothesize that the pre-training dataset used by the original BERT model already has sufficient toxic and non-toxic texts to generalize the learning encoded in the contextual embeddings.

Finding 3: Contextual embeddings do not have sufficient capacity to identify the individual toxic components of a post even with fine-tuning. For Tasks D and E, we observe that the majority of PCs for all three models perform suboptimally across all topics (0.59 to 0.76). With respect to Model 1, this observation is not surprising since contextual embeddings from the original BERT model is not pre-trained to identify individual toxic components in texts. However, with respect to Models 2 and 3, this observation is concerning, especially in instances where their PCs underperform the benchmark, since it suggests that there may be limitations for contextual embeddings to encode additional information on toxicity through fine-tuning with post-level toxicity and/or rationales for certain topics. It may also be the case that our fine-tuning approach is inadequate for these two relatively more challenging tasks compared to Tasks A, B and C.

Finding 4: The performance impact from fine-tuning depends on tasks, topics and data used in fine-tuning. We observe that fine-tuning has relatively little impact on performance (+/- 0.00) across all topics for Tasks A, B and C. This observation suggests that contextual embeddings from the original BERT model already encodes information needed to identify mentions of topics regardless of toxicity. On the other hand, fine-tuning is much more impactful across all topics for Tasks D and E with performance boost as high as +0.08, which is consistent with our previous finding that the contextual embeddings from the original BERT model may not sufficiently encode information necessary for these tasks where fine-tuning is able to help encode additional new information.
We observe that fine-tuning does not yield consistent performance changes in the same direction. For example, for Task D, Models 2 and 3 yield performance improvement across all topics, with the most and the least for profanity mentions (+0.07) and Jewish mentions (+0.01), respectively. We hypothesize that the performance gap has to do with the amount of information relevant to the task encoded during the pre-training of the original BERT model for each topic where there is perhaps less information on profanity than on Jewish, which makes fine-tuning more impactful on the topic of profanity on this specific task.

There are also instances where fine-tuning yields opposite results. For example, on the topic of Jewish for Task C, Model 2's performance degrades (-0.01) while Model 3's performance improves (+0.01). This observation is interesting because it suggests that fine-tuning with post-level toxicity only impedes the ability of the contextual embeddings to encode information for this task, but that fine-tuning with both post-level toxicity and rationales together improves such ability. We hypothesize that the performance results from fine-tuning may vary for different tasks and topics based on the combinations of fine-tuning objectives. Future work is warranted to gauge the performance impact from using different combinations of fine-tuning approaches (e.g. using rationale only).

Finding 5: Fine-tuning may allow new information to be encoded in the earlier hidden layers of the Transformer stack. Since we conduct our probing experiments using contextual embeddings from all layers across all three models, we are able to generate interesting insights into how each layer's performance changes based on fine-tuning with post-level toxicity and rationales. In general, we expect the most performance gains coming from the final hidden layers of the Transformer stack due to their proximity to the classification layer. However, our observations reveal otherwise for certain instances.

For example, for Task E, the highest layer-wise performance boosts in Models 2 and 3 are found in the zeroth layer (i.e. embedding layer) for Islam and LGBTQ mentions, respectively. This observation suggests that different fine-tuning approaches may enable the models to propagate new learnings to contextual embeddings to earlier hidden layers. We hypothesize that the hidden layers with the most performance boosts from fine-tuning may contain contextual information that is the most relevant to the

tasks and the topics at hand. Again, future work is warranted to better understand the impact of layer-wise performance using different combinations of fine-tuning approaches.

5 Conclusion

We leverage the probing tasks to understand the impact from fine-tuning on a BERT model used for toxic speech detection. More concretely, we investigate the ability of contextual embeddings from the original BERT model and two fine-tuned BERT models in encoding information relevant to toxic speech detection through five distinct probing tasks covering six different topics found in the HateXplain dataset.

We arrive at several findings based on the results from our experiments. First, contextual embeddings from the original BERT model have sufficient capacity to identify mentions of topics. Second, the presence and absence of toxicity do not affect the ability of contextual embeddings in encoding information necessary to discern mentions of topics. Third, contextual embeddings from the original BERT model do not have sufficient information to distinguish the individual toxic components of a post even with fine-tuning. Fourth, performance impact from fine-tuning is not consistent and is dependent on the tasks and topics in question. Last, fine-tuning may allow models to push their newly gained information deeper into the Transformer stack to the earlier hidden layers, achieving better performance at a task.

We release our data processing and model code, and hope that our work may serve as a demonstration on using simple linear probes in understanding the impact of fine-tuning on neural networks with respect to toxic speech detection.

6 Acknowledgement

We thank Professor Mark Butler for the helpful discussions and inspirations throughout this work.

References

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. In The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China. ACM, NewYork,NY, USA, 10 pages.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. ICWSM.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the International AAAI Conference on Web and Social Media, volume 12.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection". Accepted at AAAI 2021.

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.

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the predictions of any classifier. In SIGKDD, pages 1135–1144, San Francisco, USA.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In International Conference on Learning Representations.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendix

A.1 Error Analysis

Using Models 1 and 2, we generate predictions on our test set and conduct our error analysis on the 50 false positives with the highest probability to be toxic and the 50 false negatives with the lowest probability to be toxic. We first analyze the false positives and negatives for each model independently and then examine the ones with the largest changes in the predicted probability in either direction after fine-tuning.

A.1.1 Model 1: False positives

Among the false positives generated from Model 1, we observe that mentions of certain themes using neutral and/or derogatory terms may push the model's prediction towards toxicity. The most prevalent themes are race, ethnicity, religion, violence, politics and profanity as summarized in Table A1.1. We also observe that certain words may have been misconstrued as mentions of themes that are potentially associated with toxicity in the following examples:

Example #1:

Text: "i do not blame orlando pie rats for loosing like this they are drunk from that black label 😂 😂 😂 mtn 8 ss diski wafa wafa"

Hypothesis: The model makes the incorrect prediction based on the word "black" in the context of race instead of an alcohol brand from the noun phrase "black label".

Example #2:

Text: "this dude sits around all day trying to argue with anglin others then probably bitched about targeted harassment all day"

Hypothesis: The model makes the incorrect prediction based on the word "bitched" by interpreting the token "bitch" as a noun for a derogatory term for women instead of a verb for expressing displeasure.

A.1.2 Model 1: False negatives

Among the false positives from Model 1, we observe that the occurrences of tokens "<user>" and/or "<number>" may drive the model's prediction towards non-toxicity despite the existence of readily identifiable derogatory terms. In fact, all 50 false positives contain one or more of these tokens. We therefore hypothesize that the existence of these tokens in texts is a major contributing factor of the model's decision of non-toxicity. In other words, if a text has one or more of these two tokens, then the model would predict a low probability of toxicity.

Theme	Neutral and/or derogatory terms
Race/ ethnicity	caucasian, white, black, asian, native american, chinese, white trash, nigger, ghetto, chinaman, spic, jap, mick, kraut, porch monkey, sand nigger, moon cricket
Religion	jew, jewish, christianity, christ, islam, islamic, muslim, moslem, muzzies, kike, heeb
Violence	kill, shoot, fight, nuke, pistol, nuclear, war, terrorism, violence, rape, beat, murder, holocaust
Politics	nazi, neonazi, stalin, capitalism, nationalist, democracy, democrats, republicans, liberal, trump, right, left, potus, senate
Profanity	fuck, ass shithole, mofuckas

Table A.1.1: Summary of the top five most prevalent mentions by themes among false positives from Model 1

Theme	Neutral and/or derogatory terms
Race/ ethnicity	white, black, welsh, american, western, aglo saxon, chinese, nigger, ching chong, nigress, spic, jap, mick, kraut, porch monkey, sand nigger, moon cricket, ghetto, beaner, coons, uncle tom, negress
Religion	jew, moslem, muslim, islam, christ, muzzie, kike
Politics	Conservatives, democracy, nazi, republicans, hitler, neonazis, nazi, trump, national, politicians, president
Physical/ mental conditions	retard, whale, elderly, diseased
Sexual orientation	gay, faggot, savoury ducks, dyke

Table A.1.3: Summary of the top five most prevalent mentions by themes among false positives from Model 2

We also observe that the model is unable to identity toxicity in texts containing readily identifiable derogatory terms such as "nigger", "faggot", "spic", "gook", "muzzies" and "kike". We hypothesize that the original BERT model may not be adequately exposed to some of the derogatory terms used in the dataset during pre-training. It may also be the case that the toxicity from such readily identifiable derogatory terms is drawn out by the non-toxicity from other neutral terms in the texts.

A.1.3 Model 2: False positives

Among the false positives from Model 2, we observe the same pattern as Model 1 where mentions of certain

themes using neutral and/or derogatory terms may push the model's prediction towards toxicity. For this model, the most prevalent themes are race, ethnicity, religion, politics, physical and mental conditions and sexual orientation as summarized in Table A1.3.

A.1.4 Model 2: False negatives

Among the false negatives from Model 2, we again observe the same patterns as Model 1. We observe that the occurrences of tokens "<user>" and/or "<number>" may drive the model's prediction towards non-toxicity in spite of the existence of readily identifiable derogatory terms. However, it is worth noting that the model seems to be less reliant on these tokens as a shortcut for non-toxicity compared to Model 1 given that only 19 and 9 of the false negatives contain the tokens <user> and <number>, respectively. We also observe that the model may have trouble recognizing certain derogatory terms such as "nigga", "whore", "bitch", "ass", "hoe", "ghetto", "dyke" and "redneck".

We observe that identifying toxicity in some posts may involve deciphering the author's prejudices, which may be a challenging task for the model. For example:

Text: "<user> <user> <user> absolutely right man legal immigrants who work hard and and are law abiding citizens are not the problem the context of my comment was to highlight the devastating effects of refugee crisis post <number> that has plagued eu and canada with forced immigrants from uncivilized society"
Rationale: The author has a prejudice towards immigrants in that they come from uncivilized society.

We hypothesize that the model is unable to tease out the author's prejudices in order to make the correct determination of the post's toxicity.

We also observe that identifying toxicity may require background information that is not available in the posts. Without adequate background information, the model may not be able to identify toxicity correctly. For example:

Text: "<user> i mean if sexual harassment is just mean girl drama"
Rationale: The author equates sexual harassment with the drama as seen from Mean Girls the movie.

In this post, the writer attempts to minimize the severity of sexual harassment with this comparison. We hypothesize that our model fails to identify toxicity in this post because it does not know anything about Mean Girls the movie and therefore does not know what "mean girl drama" means to make a correct prediction.

A.1.5 Model 1 vs. Model 2: Error analysis

In this section, we take our error analysis a step further. We take this opportunity to compare the predictions from Models 1 and 2. More concretely, we compare the predicted probability to be toxic for all the posts in our test set and review the ones with the largest changes in the predicted probability in either direction. Given that Model 2 achieves better test metrics compared to Model 1, the assumption is that Model 2 is better than the Model 1 in identifying posts that are toxic.

A.1.5.1 Less toxicity after fine-tuning

We review the test posts for which Model 2 predicts a lower probability to be toxic than Model 1. In other words, fine-tuning leads to a lower predicted probability to be toxic for these test posts.

We observe that Model 2 no longer automatically deems the mentions of certain themes using neutral terms as toxic in the following examples:

- "wow so that lil black girl was lying about them white boys cutting her dreads off"
- "yeah it a sad for in america when white supremacists start a violence protest"
- "of all the women coming forward lately to accuse prominent men of sexual harassment or rape this woman has to be the bravest thus far"
- "washed my all white vans for my holiday they have gone yellow lol i want to cry"
- "shout out to my other queer babes with trauma struggling with expressing your sexuality is not homophobic 😘🥰😘 k love u all"

We observe that, despite the use of derogatory terms, Model 2 appears to be better equipped to consider the context in which such terms are used while determining whether a text is toxic in the following examples:

- "niggas coppin designer glocks now imagine a nigga wit no aim having enough money to buy a designer pistol"
- "thigh high black boots burberry all day bitches"
- "a nigga got super drunk and walked home not drinkin till vegas lol"
- "it ain t even noon yet and niggas lying"
- "regression is a bitch i haven't been this angry in a long time"

We observe that, in the absence of any mentions of themes that may potentially be toxic, Model 2 appears to be better at predicting such posts as non-toxic in the following examples:

- "shout out to whoever invented hot sauce"
- "that thing where you tie one running shoe than get distracted before shoe number two"
- "🖤 whoever lectures criminal procedure deserves happiness life and wealth my week just became great"
- "rt <user> those who hate you for being real usually love your mask"
- "real dysphoria hours why are my arms so hairy"

Based on these observations, it appears that fine-tuning allows the model to avoid taking shortcuts (e.g. automatically marking a post as toxic based on certain words) and better recognize the context in which a word is used (e.g. use of profanity in expressions).

A.1.5.2 More toxicity after fine-tuning

We review the test posts for which Model 2 predicts a higher probability to be toxic than Model 1. In other words, fine-tuning leads to a higher predicted probability to be toxic for these test posts.

We observe that Model 2 is better at recognizing mentions of derogatory terms, including ones consisting of multiple words, in posts while making the determination of a text's toxicity, in the following examples

- "<user> new look ching chong"
- "<user> <user> they're speaking retard jaja"
- "<user> tonight i ate kike <time>"
- "<user> in every country where muzzies are in minority the want minority rights in every country where muzzie are majority there are no minority rights"
- "what faggot curates live topics"

While fine-tuning does lead to more correct predictions, it also yields some incorrect predictions:

- "<user> respect the based nigress 😤 he would be nothing without her"
- "<user> i love faggots and i love being queer i wear my rainbow stripes loud and motherfucking proud 🌈"
- "<user> words are important af and change with the times we no longer describe people as retarded faggot or even indian it called progress forcing people who want things from government to treat others who also pay taxes respectfully seems ok to me"
- "real nigger me please"
- "baby was a black sheep baby was a whore baby gotten big and baby getting bigger baby got some then baby want more baby baby baby was a rock roll nigger"

We also notice that Model 2 has trouble correctly predicting a post's toxicity when neutral terms associated with religion and sexual orientation are used, such as the following:

- "<user> queer jews unite 💘 🥰"
- "<user> " lets go gays ""

Based on these observations, it appears that while fine-tuning helps the model in identifying toxic posts based on certain derogatory terms, it also leads to incorrect predictions especially when neutral and/or derogatory terms associated with certain themes are used in a non-toxic manner.

A.2 LIME coefficients

Using LIME, we extract the coefficients of the learned sparse linear model fitted on the input perturbations for both models. A word with a positive coefficient means that repeating the same word more times in the post moves the prediction towards toxicity. Conversely, a word with a negative coefficient means that repeating the same word more times in the post moves the prediction towards non-toxicity.

A.2.1 Model 1: LIME coefficients

In this section, we provide a summary based on our review of the LIME coefficients as generated using the Model 1. The top 10 and bottom 10 LIME coefficients are summarized in Tables A.2.1(a) and A.2.1(b), respectively. The following are our observations:

- The coefficients range from -0.1576 ("topics" from "what faggot curates live topics") to 0.2199 ("fuck" from "fuck off retard").
- Toxic terms that are readily identifiable by a human reader have high positive coefficients. Examples include profanity (e.g. "fuck") and derogatory terms towards specific groups (e.g. "retard", "slut", "bitch", "goatfucker", "faggot", "mudshark", "sheboon", "muzrat", "redneck", "beaners").
- When used in conjunction with readily identifiable toxic terms, articles (e.g. "a" from "a redneck woman"), pronouns (e.g. "you" in "you niggers are dumb") and verbs (e.g. "are" in "white people are fucking retarded") also have relatively high positive coefficients.
- When used in non-toxic context as deemed by annotators, certain words have relatively high positive coefficients (e.g. "jews" in "just call them jews", "moslem" in "gay vs moslem")

Word	Post	Coefficient
fuck	fuck off retard	0.2199
fucking	fucking slant retard	0.2024
nail	chipped black nail polish	0.1980
niggershitbullshit	i do not listen to niggershitbullshit	0.1778
jews	just call them jews	0.1757
you	you niggers are dumb	0.1698
fuck	fuck the nigger bitches	0.1688
retard	the blur the retard	0.1685
niglets	those are called niglets	0.1678
massive	massive muzrat espionage	0.1676

Table A.2.1(a): 10 words from test posts with the highest LIME coefficient from Model 1

Word	Post	Coefficient
user	<user> must be free gold retarded people	-0.1043
user	<user> oh my the niggers gonna be pissed	-0.1066
user	<user> <user> jews are white now	-0.1116
user	<user> lmao 😂 eses charged me under <money> fuck them white folk lmao	-0.1151
user	<user> <user> ching chong <number>	-0.1163
user	<user> omg yes arab muslims rise 👁️ 👄 👁️	-0.1233
user	<user> aw it ' retarded	-0.1265
choice	teen choice blackpink choice international artist <user>	-0.1315
user	<user> " lets go gays "	-0.1451
topics	what faggot curates live topics	-0.1576

Table A.2.1(b): 10 words from test posts with the lowest LIME coefficient from Model 1

- The token "<user>" has the highest negative coefficients across almost all test samples. The mere existence of the token in the text appears to push the predictions towards non-toxic in spite of the existence of readily identifiable toxic terms (e.g. "<user> <user> ur fuckin retarded").
- Readily identifiable toxic terms with multiple words (e.g. "ching chong") have coefficients close to zero.

4.2.2 Model 2: LIME coefficients

In this section, we provide a summary based on our review of the LIME coefficients as generated using Model 2. The top 10 and bottom 10 LIME coefficients are summarized in Tables A.2.2(a) and A.2.2(b), respectively. The following are our observations:

- The coefficients range from -0.3168 ("kenyun" from "kenyun paki at <number>") to 0.9097 ("faggots" from "sit down faggots").
- The terms with high positive coefficients appear to be dominated by derogatory terms towards specific groups (e.g. "faggots", "goatfucker", "sheboon", "kike", "retard", "mudshark", "nigger", "niglets", "beaners", "nigress", "coons", "muzzies", "bitches", "muzrat", "mudslimes", "spic", "dykes").
- Certain terms used in non-toxic context as deemed by annotators have relatively high coefficients (e.g. "jews" in "just call them jews", "homosexuals" in "how do you see or treat homosexuals", "gays" in "<user> lets go gays").
- The terms with high negative coefficients are quite

Word	Post	Coefficient
faggots	sit down faggots	0.9097
goatfucker	stealth goatfucker <number> 😂 😂 😂	0.8955
sheboon	keep waiting sheboon	0.8937
kike	this kike is still alive wtf	0.8920
kike	was engels a kike	0.8895
kike	they are going to figure out a way to kike bitcoin	0.8788
kike	enjoy your kike faith	0.8742
kike	do not break your arm jerking yourself off on twitter kike	0.8742
faggot	unrelated <user> is a faggot	0.8734
kike	more substance than anything you are ever going to post you disinfo kike	0.8724

Table A.2.2(a): 10 words from test posts with the highest LIME coefficient from Model 2

Word	Post	Coefficient
calling	loll calling me a zhid and pedophile when i am neither	-0.1745
nigga	loved 7 0 comedy you could say nigga paki and be lazy and chase women with big bristols and john and larry were the brighton element	-0.1748
stani	another bbc poster paki stani	-0.1762
thanks	thanks jews how it started	-0.1844
me	<user> nigga fucking called me a faggot when i said these things to him	-0.1900
white	<user> <user> jews are white now	-0.1930
called	<user> nigga fucking called me a faggot when i said these things to him	-0.2046
called	<user> dis nigga called us coons and niggers with a er am i missing something lmao	-0.2183
crayon	and they could not even get attractive models to wear those retarded crayon scratching designs	-0.2696
kenyun	kenyun paki at <number>	-0.3168

Table A.2.2(b): 10 words from test posts with the lowest LIME coefficient from Model 2

diverse and appear to be non-toxic when used on a standalone basis (e.g. "calling", "thanks", "me").
- Readily identifiable toxic terms with multiple words (e.g. "ching chong") have non-zero coefficients.

A.2.3 Model 1 vs. Model 2: LIME coefficients

To better understand the impact of fine-tuning, we proceed to compare the LIME coefficients as generated from Models 1 and 2. More concretely, for each word in each post from the test set, we compute the difference in LIME coefficients from the two models. The rationale for this comparison is that the difference in LIME coefficients may serve as a proxy for each word's toxicity before and after fine-tuning. The top 10 and bottom 10 words with the highest increases and decreases in LIME coefficients are summarized in Tables A.2.3(a) and A.2.3(b), respectively.

We observe dramatic increases in coefficients of readily identifiable derogatory terms such as "kike", "retard", "nigress", "faggot", "negress", "sheboon", "coon", "goatfucker", "muzzies", "beaners", "nigger", "faggotry", "spic", "mudshark", after fine-tuning. The most increase is found in the word "kike" from the post "more substance than anything you are ever going to post you disinfo kike" from 0.0426 to 1.0000. This observation is consistent with our error analysis that Model 2 is better equipped to recognize mentions of derogatory terms than Model 1.

We observe relatively moderate decreases in coefficients of neutral terms such as "kenyun", "crayon", "called", "thanks", "nail", "me", "i", "and", "them", which signals that fine-tuning allows the model to better distinguish between neutral and derogatory terms. For example, the coefficient of the word "non" from the post "non lesbians stop saying dyke" drops from 0.0507 to -0.1004. This observation is also consistent with our error analysis that Model 2 no longer automatically deems neutral terms as toxic.

We also notice decreases in coefficients of readily identifiable derogatory terms used in a non-toxic manner. For example, the coefficient of the word "nigga" from the post "calling someone nigga is acceptable but calling someone faggot it not the same" drops from 0.0358 to -0.1490. This observation aligns with our error analysis that fine-tuning allows the model to better understand the context in which readily identifiable derogatory terms are used in a non-toxic manner.

Word	Post	Change in coefficient
kike	more substance than anything you are ever going to post you disinfo kike	+0.8298
kike	solo que soy de derecha a l <percent> el virus kike no va con jesús	+0.8263
kike	i m glad ds is back on dailystormer top and no kike will ever hold our movement down	+0.8226
kike	i am happy once the last kike hangs from a tree	+0.8197
kike	imagine unironically believing the high iq kike meme	+0.8138
kike	this kike is still alive wtf	+0.8105
kike	hey how about you go be a kike somewhere else schlomo sound good	+0.8097
kike	they are going to figure out a way to kike bitcoin	+0.8066
kike	ya you really hurt my feelings what you had has a profound effect on me you must be a kike to think like this	+0.8053
kike	do not break your arm jerking yourself off on twitter kike	+0.8052

Table A.2.3(a): 10 words from test posts with the highest increase in LIME coefficient between Models 1 and 2

Word	Post	Change in coefficient
called	<user> nigga fucking called me a faggot when i said these things to him	-0.1845
nigga	calling someone nigga is acceptable but calling someone faggot it not the same	-0.1848
me	do not call me a nigger i am white	-0.1848
nail	chipped black nail polish	-0.1863
nigga	loved 7 0 comedy you could say nigga paki and be lazy and chase women with big bristols and john and larry were the brighton element	-0.1916
thanks	thanks jews how it started	-0.1942
stani	another bbc poster paki stani	-0.2244
called	<user> dis nigga called us coons and niggers with a er am i missing something lmao	-0.2255
crayon	and they could not even get attractive models to wear those retarded crayon scratching designs	-0.2809
kenyun	kenyun paki at <number>	-0.3644

Table A.2.2(b): 10 words from test posts with the highest decrease in LIME coefficient between Models 1 and 2

A.3 Summary of experimental results

We provide a summary of the results from our probing experiments in Tables A.3(a) to A.3(e) for Tasks A to E, respectively.

	Task 1A: Profanity					Task 2A: Women					Task 3A: LGBTQ					Task 4A: African					Task 5A: Jewish					Task 6A: Islam				
Layer	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3
12	0.98	0.98	0.99	-0.01	0.00	0.97	0.95	0.96	-0.02	0.00	0.98	0.95	0.98	-0.02	0.00	0.98	0.95	0.97	-0.02	-0.01	0.97	0.95	0.98	-0.02	0.01	0.97	0.96	0.98	-0.01	0.01
11	0.98	0.98	0.98	0.00	0.00	0.97	0.95	0.97	-0.01	0.00	0.98	0.96	0.98	-0.02	0.00	0.98	0.97	0.98	0.00	0.00	0.98	0.96	0.98	-0.02	0.00	0.98	0.97	0.99	-0.01	0.01
10	0.98	0.98	0.99	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.98	0.97	0.99	-0.01	0.01	0.98	0.98	0.98	0.00	0.00	0.98	0.97	0.98	-0.01	0.00	0.98	0.98	0.99	0.00	0.01
9	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.98	0.00	0.01	0.98	0.96	0.98	-0.01	0.01	0.98	0.98	0.98	-0.01	0.00	0.98	0.97	0.98	-0.01	0.00	0.98	0.97	0.98	0.00	0.00
8	0.99	0.99	0.99	0.00	0.00	0.97	0.98	0.98	0.01	0.01	0.98	0.97	0.98	-0.01	0.00	0.99	0.98	0.98	0.00	0.00	0.98	0.97	0.98	-0.01	0.00	0.99	0.98	0.98	-0.01	0.00
7	0.99	0.99	0.99	0.00	0.00	0.98	0.97	0.98	0.00	0.00	0.98	0.97	0.99	-0.01	0.00	0.99	0.98	0.98	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.98	0.98	0.00	0.00
6	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.98	0.00	0.00	0.98	0.98	0.99	-0.01	0.00	0.98	0.99	0.98	0.00	0.00	0.98	0.98	0.98	-0.01	0.00	0.99	0.98	0.98	-0.01	-0.01
5	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.98	0.00	0.00	0.99	0.98	0.99	-0.01	0.00	0.99	0.99	0.98	0.00	0.00	0.98	0.98	0.98	-0.01	-0.01	0.99	0.98	0.99	-0.01	0.00
4	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.98	0.99	0.00	0.00	0.98	0.97	0.97	-0.01	-0.01	0.99	0.99	0.99	0.00	0.00
3	1.00	0.99	0.99	0.00	0.00	0.97	0.98	0.97	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00
2	0.99	0.99	0.99	0.00	0.00	0.98	0.97	0.98	-0.01	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00
1	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00
0	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.98	0.98	0.98	0.00	0.00
Avg	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.98	0.00	0.00	0.98	0.98	0.98	-0.01	0.00	0.99	0.98	0.98	0.00	0.00	0.98	0.97	0.98	-0.01	0.00	0.98	0.98	0.99	0.00	0.00

Table A.3(a). Performance of Task A by topics as measured by average F1 score achieved by PCs trained with contextual embeddings across all layers from the three models ("M#"). Δ2 and Δ3 denote performance differences between M2 and M1 and between M3 and M1, respectively.

	Task 1B: Profanity					Task 2B: Women					Task 3B: LGBTQ					Task 4B: African					Task 5B: Jewish					Task 6B: Islam				
Layer	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3
12	0.98	0.97	0.98	0.00	0.00	0.95	0.94	0.97	-0.02	0.01	0.97	0.94	0.98	-0.03	0.01	0.98	0.95	0.97	-0.02	-0.01	0.97	0.94	0.98	-0.02	0.01	0.97	0.96	0.97	-0.01	0.00
11	0.99	0.98	0.99	-0.01	0.00	0.96	0.96	0.97	0.00	0.01	0.97	0.95	0.98	-0.02	0.01	0.98	0.97	0.98	-0.01	0.00	0.97	0.96	0.98	-0.01	0.00	0.98	0.97	0.98	-0.01	0.00
10	0.99	0.98	0.99	-0.01	0.00	0.96	0.96	0.97	0.00	0.01	0.97	0.95	0.98	-0.02	0.01	0.97	0.97	0.98	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.98	0.97	0.98	-0.01	0.00
9	0.99	0.98	0.99	-0.01	0.00	0.96	0.96	0.98	0.00	0.01	0.97	0.95	0.98	-0.02	0.01	0.98	0.97	0.98	0.00	0.01	0.97	0.96	0.97	-0.01	-0.01	0.98	0.97	0.97	-0.01	-0.01
8	0.99	0.99	0.99	0.00	0.00	0.96	0.96	0.98	0.00	0.01	0.97	0.97	0.98	0.00	0.01	0.98	0.98	0.98	0.00	0.00	0.98	0.97	0.97	-0.01	-0.01	0.98	0.97	0.98	-0.01	-0.01
7	0.99	0.99	0.99	0.00	0.00	0.96	0.97	0.97	0.01	0.01	0.98	0.97	0.98	-0.01	0.00	0.98	0.98	0.98	0.00	0.00	0.97	0.97	0.97	0.00	0.00	0.98	0.98	0.98	-0.01	0.00
6	0.99	0.99	1.00	0.00	0.00	0.96	0.96	0.97	0.00	0.01	0.98	0.97	0.98	-0.01	0.00	0.98	0.98	0.98	0.00	0.00	0.97	0.97	0.97	-0.01	0.00	0.99	0.98	0.98	-0.01	-0.01
5	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.97	0.00	0.01	0.99	0.98	0.98	-0.01	0.00	0.99	0.99	0.98	0.00	0.00	0.98	0.97	0.97	0.00	-0.01	0.98	0.97	0.98	-0.01	0.00
4	0.99	1.00	0.99	0.01	0.00	0.97	0.97	0.97	0.00	0.00	0.99	0.98	0.98	-0.01	0.00	0.99	0.98	0.99	0.00	0.00	0.98	0.97	0.98	0.00	0.00	0.99	0.98	0.98	0.00	-0.01
3	0.99	0.99	0.99	0.00	0.00	0.97	0.96	0.98	-0.01	0.01	0.98	0.98	0.99	0.00	0.00	0.99	0.98	0.98	0.00	-0.01	0.98	0.98	0.98	-0.01	-0.01	0.99	0.98	0.98	-0.01	-0.01
2	0.99	0.99	0.99	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.99	0.98	0.99	0.00	0.00	0.98	0.98	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.98	0.99	0.98	0.00	0.00
1	0.99	0.99	0.99	0.00	0.00	0.97	0.97	0.97	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.97	0.98	0.97	0.01	0.00	0.98	0.98	0.98	-0.01	0.00
0	0.99	0.99	0.99	0.00	0.00	0.96	0.97	0.98	0.01	0.01	0.98	0.99	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.97	0.97	-0.01	-0.01	0.97	0.97	0.97	0.00	0.00
Avg	0.99	0.99	0.99	0.00	0.00	0.96	0.96	0.97	0.00	0.01	0.98	0.97	0.98	-0.01	0.00	0.98	0.98	0.98	0.00	0.00	0.97	0.97	0.97	-0.01	0.00	0.98	0.97	0.98	-0.01	0.00

Table A.3(b). Performance of Task B by topics as measured by average F1 score achieved by PCs trained with contextual embeddings across all layers from the three models ("M#"). Δ2 and Δ3 denote performance differences between M2 and M1 and between M3 and M1, respectively.

	Task 1C: Profanity					Task 2C: Women					Task 3C: LGBTQ					Task 4C: African					Task 5C: Jewish					Task 6C: Islam				
Layer	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3
12	0.97	0.98	0.98	0.01	0.02	0.96	0.94	0.96	-0.02	0.00	0.97	0.95	0.98	-0.02	0.01	0.97	0.95	0.98	-0.03	0.00	0.95	0.92	0.98	-0.04	0.02	0.96	0.95	0.98	-0.02	0.02
11	0.99	0.99	0.98	0.00	0.00	0.96	0.95	0.97	-0.01	0.01	0.97	0.96	0.99	-0.01	0.02	0.97	0.97	0.98	0.00	0.01	0.96	0.95	0.98	-0.02	0.02	0.97	0.96	0.98	0.00	0.02
10	0.98	0.99	0.98	0.01	0.00	0.96	0.96	0.96	0.00	0.00	0.98	0.96	0.98	-0.01	0.01	0.97	0.97	0.97	0.00	0.00	0.94	0.94	0.97	-0.01	0.03	0.97	0.96	0.99	0.00	0.02
9	0.98	0.99	0.99	0.02	0.02	0.96	0.96	0.97	0.00	0.01	0.98	0.97	0.99	-0.01	0.01	0.97	0.97	0.98	0.00	0.01	0.96	0.95	0.97	-0.01	0.00	0.97	0.96	0.98	-0.01	0.01
8	0.98	0.99	0.99	0.01	0.01	0.96	0.98	0.97	0.02	0.02	0.98	0.97	0.99	-0.01	0.00	0.98	0.98	0.98	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.96	0.96	0.98	-0.01	0.01
7	0.98	0.98	0.99	0.01	0.02	0.96	0.97	0.97	0.01	0.01	0.97	0.97	0.99	0.00	0.01	0.98	0.98	0.98	0.00	0.01	0.96	0.96	0.98	0.00	0.02	0.98	0.97	0.97	-0.01	0.00
6	0.99	0.99	0.99	0.00	0.01	0.96	0.96	0.97	0.00	0.00	0.97	0.98	0.99	0.00	0.01	0.97	0.98	0.98	0.00	0.01	0.96	0.96	0.96	0.00	0.00	0.98	0.97	0.98	-0.01	0.00
5	0.98	0.99	1.00	0.01	0.01	0.97	0.97	0.97	0.00	0.00	0.98	0.98	0.98	-0.01	-0.01	0.98	0.98	0.98	0.00	0.00	0.96	0.95	0.96	-0.01	0.00	0.98	0.98	0.98	0.00	0.00
4	0.99	0.99	0.98	0.00	-0.01	0.97	0.97	0.97	0.00	0.00	0.98	0.98	0.99	0.00	0.01	0.98	0.98	0.98	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.98	0.97	0.99	-0.01	0.01
3	0.98	0.99	1.00	0.00	0.01	0.97	0.97	0.97	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.98	0.95	0.97	-0.03	-0.01	0.97	0.98	0.99	0.01	0.01
2	0.99	0.99	0.99	0.00	0.00	0.98	0.98	0.98	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.99	0.98	0.98	0.00	0.00	0.96	0.96	0.95	0.00	-0.01	0.98	0.99	0.99	0.02	0.01
1	0.98	0.99	0.99	0.01	0.01	0.98	0.98	0.97	0.00	-0.01	0.99	0.99	0.99	0.00	0.01	0.99	0.99	0.98	0.00	0.00	0.96	0.95	0.97	-0.01	0.01	0.98	0.98	0.98	0.00	0.00
0	0.99	0.99	0.99	0.00	0.00	0.97	0.98	0.97	0.01	0.00	0.99	0.99	0.99	0.00	0.01	0.98	0.98	0.99	0.00	0.00	0.97	0.96	0.97	-0.01	0.00	0.99	0.98	0.97	-0.01	-0.01
Avg	0.98	0.99	0.99	0.01	0.01	0.97	0.97	0.97	0.00	0.00	0.98	0.97	0.99	-0.01	0.01	0.98	0.97	0.98	0.00	0.00	0.96	0.95	0.97	-0.01	0.01	0.97	0.97	0.98	0.00	0.01

Table A.3(c). Performance of Task C by topics as measured by average F1 score achieved by PCs trained with contextual embeddings across all layers from the three models ("M#"). Δ2 and Δ3 denote performance differences between M2 and M1 and between M3 and M1, respectively.

	Task 1D: Profanity					Task 2D: Women					Task 3D: LGBTQ					Task 4D: African					Task 5D: Jewish					Task 6D: Islam				
Layer	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3
12	0.67	0.80	0.76	0.13	0.09	0.66	0.68	0.70	0.02	0.04	0.67	0.78	0.72	**0.11**	**0.05**	0.73	0.74	0.76	0.01	0.04	0.72	0.74	0.77	**0.02**	**0.04**	0.69	0.79	0.77	**0.09**	0.08
11	0.63	0.83	0.79	**0.19**	**0.16**	0.66	0.68	0.68	0.02	0.02	0.70	0.77	0.74	0.07	0.03	0.71	0.78	0.75	**0.07**	0.04	0.72	0.74	0.71	**0.02**	0.00	0.67	0.76	0.77	**0.09**	**0.10**
10	0.63	0.79	0.78	0.16	0.15	0.66	0.68	0.69	0.03	0.04	0.67	0.75	0.72	0.08	**0.05**	0.70	0.76	0.77	0.06	**0.07**	0.73	0.71	0.75	-0.02	0.01	0.70	0.75	0.73	0.05	0.04
9	0.64	0.75	0.76	0.12	0.13	0.64	0.67	0.69	**0.04**	0.06	0.68	0.76	0.73	0.08	**0.05**	0.72	0.75	0.76	0.03	0.03	0.72	0.73	0.71	0.01	-0.01	0.71	0.72	0.75	0.01	0.05
8	0.65	0.73	0.75	0.09	0.11	0.61	0.65	0.71	**0.04**	**0.09**	0.70	0.72	0.70	0.02	0.00	0.70	0.74	0.73	0.04	0.03	0.73	0.71	0.72	-0.02	-0.01	0.68	0.72	0.75	0.03	0.06
7	0.63	0.68	0.72	0.06	0.09	0.64	0.64	0.70	0.00	0.06	0.72	0.73	0.71	0.01	0.00	0.72	0.73	0.74	0.01	0.01	0.70	0.72	0.71	**0.02**	0.01	0.69	0.71	0.73	0.01	0.04
6	0.62	0.70	0.74	0.07	0.11	0.64	0.65	0.67	0.02	0.03	0.69	0.73	0.72	0.04	0.03	0.71	0.71	0.73	0.00	0.02	0.70	0.72	0.73	**0.02**	0.03	0.68	0.65	0.71	-0.04	0.03
5	0.63	0.68	0.73	0.04	0.10	0.67	0.67	0.68	-0.01	0.01	0.69	0.73	0.73	0.03	0.03	0.72	0.71	0.74	-0.01	0.02	0.68	0.69	0.71	0.01	0.03	0.67	0.71	0.73	0.04	0.05
4	0.67	0.69	0.69	0.02	0.02	0.65	0.67	0.64	0.02	-0.01	0.69	0.70	0.69	0.01	0.01	0.69	0.73	0.72	0.04	0.03	0.71	0.72	0.70	0.01	-0.01	0.68	0.67	0.71	-0.01	0.03
3	0.65	0.69	0.70	0.04	0.06	0.63	0.65	0.67	0.01	0.04	0.68	0.69	0.68	0.01	0.00	0.72	0.72	0.70	0.00	-0.02	0.72	0.74	0.75	**0.02**	0.03	0.65	0.68	0.69	0.02	0.04
2	0.66	0.69	0.69	0.03	0.03	0.67	0.65	0.69	-0.02	0.02	0.69	0.71	0.68	0.02	-0.01	0.72	0.70	0.72	-0.01	0.01	0.73	0.70	0.73	-0.03	0.00	0.65	0.65	0.68	0.00	0.02
1	0.67	0.67	0.68	0.00	0.00	0.66	0.64	0.68	-0.01	0.03	0.67	0.66	0.67	-0.01	0.00	0.68	0.71	0.69	0.03	0.00	0.72	0.73	0.71	0.01	-0.01	0.62	0.61	0.65	-0.01	0.03
0	0.55	0.53	0.52	-0.01	-0.03	0.65	0.63	0.65	-0.03	-0.01	0.66	0.64	0.65	-0.03	-0.01	0.71	0.70	0.71	-0.01	0.01	0.72	0.73	0.71	0.01	-0.01	0.64	0.61	0.63	-0.03	-0.01
Avg	0.64	0.71	0.72	0.07	0.08	0.65	0.66	0.68	0.01	0.03	0.69	0.72	0.70	0.03	0.02	0.71	0.73	0.73	0.02	0.02	0.72	0.72	0.72	0.01	0.01	0.67	0.69	0.71	0.02	0.04

Table A.3(d). Performance of Task D by topics as measured by average F1 score achieved by PCs trained with contextual embeddings across all layers from the three models ("M#"). Δ2 and Δ3 denote performance differences between M2 and M1 and between M3 and M1, respectively. Figures in bold denote highest layer-wise performance improvements.

	Task 1E: Profanity					Task 2E: Women					Task 3E: LGBTQ					Task 4E: African					Task 5E: Jewish					Task 6E: Islam				
Layer	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3	M1	M2	M3	Δ2	Δ3
12	0.62	0.63	0.63	0.00	0.01	0.67	0.69	0.71	**0.02**	0.04	0.68	0.69	0.70	0.00	0.02	0.74	0.76	0.76	0.02	0.02	0.71	0.75	0.75	**0.04**	**0.04**	0.61	0.67	0.66	**0.06**	0.05
11	0.61	0.59	0.60	-0.01	-0.01	0.69	0.70	0.71	0.01	0.02	0.66	0.71	0.72	0.05	**0.05**	0.72	0.78	0.79	0.05	**0.06**	0.74	0.75	0.74	0.01	0.00	0.62	0.66	0.68	0.03	**0.06**
10	0.60	0.63	0.60	**0.02**	0.00	0.68	0.69	0.72	0.01	0.04	0.70	0.73	0.71	0.03	0.01	0.73	0.78	0.78	0.05	0.05	0.71	0.72	0.75	0.02	**0.04**	0.62	0.65	0.67	0.03	0.05
9	0.60	0.62	0.62	**0.02**	0.02	0.70	0.72	0.73	**0.02**	0.02	0.74	0.73	0.69	-0.01	-0.05	0.72	0.77	0.77	**0.06**	**0.06**	0.72	0.73	0.74	0.02	0.03	0.63	0.66	0.66	0.03	0.03
8	0.61	0.62	0.63	0.01	0.02	0.69	0.71	0.74	**0.02**	**0.06**	0.69	0.73	0.68	0.03	-0.02	0.74	0.80	0.77	**0.06**	0.03	0.73	0.71	0.75	-0.02	0.02	0.64	0.67	0.68	0.03	0.05
7	0.57	0.58	0.61	0.01	**0.05**	0.73	0.74	0.73	0.01	0.00	0.67	0.71	0.66	0.04	-0.01	0.75	0.77	0.75	0.02	0.00	0.72	0.74	0.72	0.02	0.00	0.64	0.64	0.70	0.00	**0.06**
6	0.57	0.59	0.59	**0.02**	0.02	0.71	0.72	0.72	0.00	0.01	0.72	0.71	0.69	-0.01	-0.03	0.73	0.79	0.75	**0.06**	0.01	0.72	0.74	0.67	0.01	-0.05	0.66	0.67	0.65	0.01	-0.01
5	0.60	0.57	0.62	-0.03	0.02	0.71	0.73	0.71	**0.02**	0.01	0.65	0.72	0.70	**0.07**	**0.05**	0.73	0.74	0.76	0.01	0.03	0.69	0.72	0.72	0.02	0.02	0.65	0.66	0.66	0.01	0.01
4	0.59	0.57	0.59	-0.02	0.00	0.72	0.71	0.73	-0.01	0.01	0.69	0.69	0.71	0.00	0.02	0.75	0.76	0.74	0.02	0.00	0.70	0.73	0.69	0.03	0.00	0.66	0.66	0.62	0.00	-0.04
3	0.56	0.57	0.58	0.01	0.02	0.72	0.74	0.72	**0.02**	0.00	0.67	0.67	0.70	0.00	0.03	0.75	0.74	0.73	-0.02	-0.02	0.72	0.74	0.72	0.02	0.00	0.62	0.62	0.65	0.00	0.03
2	0.57	0.55	0.59	-0.02	0.01	0.73	0.73	0.70	-0.01	-0.04	0.66	0.68	0.70	0.02	0.03	0.74	0.73	0.72	-0.02	-0.03	0.71	0.71	0.73	0.00	0.02	0.64	0.65	0.64	0.01	0.00
1	0.58	0.58	0.61	0.00	0.03	0.71	0.72	0.72	0.01	0.01	0.68	0.66	0.65	-0.02	-0.03	0.72	0.73	0.72	0.01	0.00	0.70	0.71	0.71	0.01	0.01	0.65	0.63	0.65	-0.02	0.00
0	0.56	0.54	0.54	-0.02	-0.02	0.69	0.70	0.70	**0.02**	0.01	0.60	0.62	0.64	0.02	**0.05**	0.70	0.71	0.74	0.02	0.04	0.66	0.67	0.70	0.01	**0.04**	0.60	0.66	0.63	**0.06**	0.03
Avg	0.59	0.59	0.60	0.00	0.01	0.70	0.72	0.72	0.01	0.01	0.68	0.70	0.69	0.02	0.01	0.73	0.76	0.75	0.03	0.02	0.71	0.72	0.72	0.01	0.01	0.63	0.65	0.66	0.02	0.02

Table A.3(e). Performance of Task E by topics as measured by average F1 score achieved by PCs trained with contextual embeddings across all layers from the three models ("M#"). Δ2 and Δ3 denote performance differences between M2 and M1 and between M3 and M1, respectively. Figures in bold denote highest layer-wise performance improvements.