Text Content Moderation Model to Detect Sexually Explicit Content

Natural Language Processing

Santosh Addanki
Department of Computer Science
Stanford University
saddanki@stanford.edu

Nandana Murthy
Department of Computer Science
Stanford University
nandanam@stanford.edu

Abstract

Our objective is to build a Transformer based classification model for the recognition of Sexually Explicit Speech and do a comparative analysis with other(fine-tuned) models in the Sexually Explicit domain. Our approach was to take an open-source model for general English language understanding and pre-train it on sexually explicit data with the objective of Masked Language modelling before fine-tuning for the sexually explicit detection task.

A lack of standardized definitions of "sexually explicit" for all the available public datasets, meant we could not simply use all these datasets to train. Directly using them for model fine-tuning led to very poor performance and resulted in many false positives. We experimented with a variety of methods for incorporating “more closely related datasets” with varying degrees of success.

For Baseline, we fine-tuned an off-the-shelf BERT based model (miniLM [1]) to achieve the classification task. This is the standard training pipeline for transformer-based models and provided a good baseline for further experimentation. We experimented with several pre-trained model architectures and fine-tuning them to keep pushing the baseline. The best baseline was a fine-tuned RoBERTa [2] model with a Precision of 0.81 and F1 score of 0.74 for the sexually explicit class.

We further explored Data Augmentation and Vocabulary extensions with varying degrees of success detailed below. We finally explored domain adaptation using Masked Language Modelling.

Results: Data augmentation techniques like capitalization yielded significant impact on performance increasing precision of the sexually explicit class from 0.81 to 0.95, and F1-score from 0.74 to 0.93 over the base model.

We further explored several Vocabulary extension techniques detailed below with no success. The best performing model with vocabulary extension offered no performance lift over the baseline model, in fact it led to a deterioration in model performance with the precision dropping to 0.92 and the F1 score to 0.88, emphasizing the need to explore research in Vocabulary extension.

Pre-training with masked language modeling, starting with pre-trained RoBERTa base[3] led to a decrease in Precision to 0.90 and F1-Score of 0.61 for the sexually explicit class. However we noticed an increase in Precision to 0.96 and F1 score to 0.98 for the negative class proving the hypothesis that masked language models bring significant boost to F1-Score when we have significantly large data representing a specific class during fine-tuning.

1 Introduction

Social media platforms increasingly play a pivotal role in both spreading and combating sexually explicit content today. Detecting sexually explicit content is a unique challenge that is dynamic and evolving rapidly. Context and subtle nuances vary widely across cultures, languages, and regions.

Additionally, sexually explicit text itself isn't always explicit. Models must be able to recognize subtleties quickly and proactively. Our objective is to build a state-of-the-art sexually explicit text content classification model, to improve recognition related to sexually explicit keywords, including slurs and slogans of sexually explicit groups, which have been considered as terms of interest.

Though there are several publicly available datasets related to sexually explicit, however, one quickly apparent issue is that definitions of "sexually explicit" vary significantly, and often not even stated. As a result, the datasets cannot simply be combined or directly applied to train for our use case, since different annotation schemes could result in different labels for the same utterance. Hence, the need to arrive at a more standardized definition of sexually explicit. "Sexually Explicit" - speech that alludes to Intercourse, masturbation, porn, sex toys and genitalia,Sexual intent, nudity and lingerie and Informational statements that are sexual in nature, affectionate activities (kissing, hugging, etc.), flirting, pet names, relationship status, sexual insults and rejecting sexual advances.

To filter for the more applicable ones, a small sample of sexually explicit utterances from each dataset was qualitatively evaluated, and further validation was focused on those that were closer to the above definition. We experimented with a variety of methods for incorporating the more closely related datasets, which are detailed below under experiments for Leveraging Public Data. Furthermore, we used the remaining datasets to pre-train a model with the Masked Language modeling objective, for which the labels were not required.

The standard transformer-based training approach of starting with an open-source model trained for general language understanding and fine-tuning it for a specific task can yield good results. However, it is well known that within a language, different types of words are used in different domains and that the same word can have different meanings and connotations in different domains. Almost all pre-trained model Architectures (MiniLM[1], RoBERTa[3], XLM-RoBERTa[4], Unitary RoBERTa[5], Unitary XLM-RoBERTa[6]) tend to use more data to produce a better general understanding of the whole language, rather than one geared toward a specific domain like in this instance sexually explicit.

Following the BioBERT[7] paper, our approach was to take an open-source model for general English language understanding, such as RoBERTa, and pre-train it on sexually-explicit data before fine-tuning for the sexually explicit detection task. To address a possible shortcoming of the approach listed in the paper, we also extended the base vocabulary with a list of approximately 100 domain-specific keywords. This extended vocabulary was used for both pre-training and fine-tuning the model. Although we expected the inclusion of keywords in the vocabulary to improve model performance, one concern with this approach was that the embeddings for the new words lacked the extensive pre-training of the rest of the vocabulary.

For example, RoBERTa is pre-trained on over 160GB of corpus, which allows it to learn useful embeddings for the words in its vocabulary. Even though the rest of the pre-trained model weights can be borrowed, the embeddings for the newly added keywords were random by default, so they were detrimental to model performance. Pre-training from scratch on a similarly sized corpus was too computationally intensive, so we borrowed strength from the pre-trained embeddings. Specifically, for each added keyword, we tokenized it using the original vocabulary and initialized its embedding as the mean of its resulting tokens before pre-training on the public datasets listed above.

Our analysis throws light on how generic pre-trained models work in newer domains and their pitfalls. Our experiments prove that domain adaptation via masked language modelling brings significant performance boost to F1 score when there are large datasets available. Further improvements can be achieved by incorporating Vocabulary extensions in various ways, and our experiments reinforce the need to push the research in this direction, of being able to find efficient techniques to use pre-trained embeddings for newer Vocabulary without extensive pre-training on large corpora and help achieve state of the art in sexually explicit text classification.

2 Approach

- Our approach is to develop a model with BERT architecture to identify sexually explicit text, rather than general English language understanding. Using the specialized pretrained model as a starting point for fine-tuning, achieves better performance on identifying sexually explicit text, demonstrating the potential benefit of domain-specific pre-training and vocabulary extension.

- We fine-tuned an off-the-shelf pre-trained model architectures (MiniLM, RoBERTa, XLM-RoBERTa, Unitary RoBERTa, Unitary XLM-RoBERTa) on the below listed datasets to achieve the classification task. This is the standard training pipeline for transformer-based models, so it provides a good baseline for further experimentation.
- We looked at the tokenization of the sexually explicit utterances. Most of the sexually explicit keywords are missing from the vocabulary and not optimally tokenizing them leading to sub optimal language model performance. To solve the tokenization problem better, we look at the top 100 sexually explicit keywords by frequency across the dataset and add them to the tokenizer for better handling of tokenization. Finally extend the model token embeddings to match the length of the new tokens.
- While we are working on Vocabulary extension techniques and domain adaptation using pre-training, we also successfully implemented some Data-Augmentation techniques detailed below with varying degrees of success.

3 Experiments

- **Data**: The definition of "Sexually Explicit" varies widely or sometimes even missing across in all of the available public datasets. Direct usage of these datasets to train for our use case, was not possible as different annotation schemes could result in different labels for the similar utterance. Based on that we choose the datasets in 1 for further training and evaluation.

  Once we merge datasets, we split them 70:20:10 as training, validation and test data. On deeper analysis of the validation data, it was found that a lot of sexually explicit samples with less confidence were not sexually explicit in reality. Hence, we decided to pre-filter for all the samples(train/validation/test) with confidence score greater than 0.8 from the labelled datasets.This resulted in a smaller data sets, but definitely a less noisy one.

Dataset	**Text Source**
Jigsaw Unintended Bias in Toxicity Classification[8]	Civil Comments
PAN12 Deception Detection:Sexual Predator Identification[9]	PAN 2012
Multilingual and Multi-Aspect Hate Speech Analysis [10]	Twitter
Offensive Comments in the Brazilian Web[11]	GitHub
Profanity Keywords[12]	GitHub
Sexual Harassment[13]	GitHub
Random Chat Data	Tagged reviews with "Un-wanted Sexual behavior". 1721 positive samples

Table 1: Table of datasets and their sources.

- **Evaluation method** We use the standard accuracy metrics like precision, recall & f1 score. As the number of positive samples are lot lesser across the datasets, we have an imbalance data problem. Hence, we look at precision of the positive class (Sexually-Explicit) and f1 score of the positive class.

  Due to the natural class imbalance for this task – relatively few sentences from almost any source are sexually-explicit – more generic metrics like overall accuracy paint an incomplete picture of real performance since high accuracy can be attained by simply classifying every utterance as not sexually-explicit. Another simple approach of classifying based on the presence of keywords could have similar coverage to deep learning-based methods, but our observations while annotating keyword-filtered data show that it would result in many false positives. As such, our primary evaluation metric was the precision for the sexually-explicit class since it tracks the proportion of false positives. Additionally, because precision can be tuned modifying classification thresholds – requiring more confidence before predicting an utterance is sexually explicit increases precision, but also increases the number of false negatives – we will consider the F1-score for the sexually-explicit class as a measure of the overall performance of the model on sexually-explicit examples.

3.1 Experimental details

3.1.1 Experiment 1 - Baseline: miniLM model with all data

Train a sexually explicit classification model by fine tuning on dataset by merging all the datasets listed above, create a training, validation, test split of 70:20:10 ratio. Train the model on the mixed dataset, validate on the validation data set and evaluate the performance on test dataset.

Result: The baseline model had a positive class precision(Sexually-explicit) at 0.79 and F1 score of 0.72. This was a good baseline to experiment further with other model architectures.

3.1.2 Experiment 2 - Baseline: RoBERTa-base model with all data

Train a sexually explicit classification model by fine tuning on the same dataset as experiment 1.

Result: Roberta being a larger model did help in improving the performance metrics slightly. This is our best baseline model with precision of positive class at 0.81 and F1 Score of positive class at 0.74.

3.1.3 Experiment 3 - Replacement Strategies + Capitalization Augmentation

The sexually explicit keywords appear in different forms and are often replaced by symbols like $, 0 (a$$, $h!t) to escape the keyword filters available on various social media. We use replacement to generate sentences with variety of such sexually explicit keywords. We use the same dataset used in experiment 1. In addition to using special characters to escape keyword filters, we have also seen in various chat conversations that speakers try to use capital letters to emphasize more on the sexually explicit keywords (ASS, SHIT). Also, phrases associated with certain sexually explicit groups are usually in all caps initialisms; e.g. "WWG1WGA", a phrase associated with Qanon.

Models tended to overfit to this trend and predicted sexually explicit for innocuous messages w/ more capital letters. So, in this experiment we augmented training data from Experiment 2 with all caps version of each utterance. We also removed utterances where capitalized version existed in validation set

Result: We observed significant improvement in performance increasing the precision to 0.95 and F1-Score to 0.93 for the sexually explicit class.

3.1.4 Experiment 4 - Naive Vocabulary Extension

To solve the tokenization problem better, we look at the top 100 sexually explicit keywords by frequency across the dataset and add them to the tokenizer for better handling of tokenization. Finally extend the model token embeddings to match the length of the new tokens. The embedding values for the new tokens are randomly initialized. With these changes, we observe better tokenization of sexually explicit keywords.

Result: While the tokenization looks better, there is a considerable drop in the accuracy metrics. This might be due to the random value initialization of the new token embeddings. A better way to initialize embedding for the new words will be to take the average of the tokens' embeddings For example, embedding of "WWG1WGA" initialized as avg. of embeddings of ("WW", "G", "1", "W", "GA"). This method resulted in precision and F1-score for the sexually explicit class of 0.88 and 0.85respectively.

3.1.5 Experiment 5: Vocab Extension with mean embedding initialization

To optimize the random initialize of the embeddings, we borrowed strength from the embeddings in the original vocabulary to improve the new words' initialization. Specifically, for each added keyword, we tokenized it under the original vocabulary, and initialized its embedding as the mean of its resulting tokens.

Result: Even taking this initialization and directly fine-tuning on the dataset led to a noticeable improvement. This method resulted in precision and F1-score for the sexually-explicit class of 0.92 and 0.88, respectively.

3.1.6 Experiment 6: Further pre-training of RoBERTa-base for MLM

Large transformer models like Bert, RoBerta[3], miniLM[1] are pre-trained on tons of common language corpus like wikipedia, common crawl etc. These models have the ability to work with multiple tasks like classification, named entity recognition, question answering to name a few. These models generalize well outside of the training data domain. However, when we are working with niche domains like sexually explicit speech or medicine like BioBERT, the structure of the sentence and the words used are very different from the style we see on wikipedia.

To leverage the benefits of these models in these domains, we need the model to adapt to the domain. To do this, we run a domain adaptation task like Masked Language Modelling. Once the model is domain adapted, we see experiments which show drastic improve in downstream tasks in that specific domain. In our experiment, we start MLM with a RoBERT-base model, approximately 150K utterances were taken from the combined public datasets, and augmented with capitalized versions of them for a total of about 300k training utterances. Training was run for 2 epochs, which took just over 7 hours on GPU Notebooks with 1 GPU, and this model was saved. It was then fine-tuned on sexually explicit classification task.

Result: This resulted in degraded performance when compared to the baseline with the precision at 0.90 and F1-Score at 0.61 for the positive class. However, we saw good jump in precision and F1-score for the negative class. This explains the need to abundant data during fine-tuning for the specific class to see benefits from the pretraining phase.

3.1.7 Experiment 7: Pre-training for MLM with Vocabulary Extension

We combined both our primary approaches - extended vocabulary and pre-training and fine-tuning the model. The model performance was below the best performing model from Masked Language Modeling.

Result: We saw a significant drop in model performance compared to pre-training and fine-tuning approaches with the precision of positive class at 0.87 and the F1-score of positive class at 0.39.

3.2 Results

The results from all experiments is summarized in the table below

Experiment		Overall		sexually explicit Class	
No.	Name	$accuracy$	$macroF1score$	$precision$	$F1score$
1.	miniLM model	0.91	0.81	0.79	0.72
2.	RoBERTa-base model	0.92	0.84	0.81	0.74
3.	Roberta Replacement Strategies + Capitalization	0.93	**0.94**	**0.95**	**0.93**
4.	Roberta Naive Vocabulary Extension	0.93	0.90	0.88	0.85
5.	Roberta Vocab Extension with mean embedding	0.96	0.91	0.92	0.88
6.	Pre-training RoBERTa-base for MLM	**0.96**	0.79	0.90	0.61
7.	Pre-training RoBERTa-base for MLM + Vocab Ext.	0.95	0.68	0.87	0.39

Table 2: Summary of results from all experiments

4 Conclusion

When working with domains which are different from general English – written / spoken, large transformer models might not perform best out of the box. Our experiments proves that it is very essential to understand the domain data and perform qualitative analysis like tokenization, embedding initialization. This analysis will throw light on how generic pretrained model works in newer domains and their pitfalls. Our experiments prove that domain adaptation via vocabulary extension and masked language modelling can bring performance boost to f1 score achieving state of the art in sexually

explicit text classification. We need to have abundant data during fine-tuning for the specific class to see benefits from the pretraining phase. The experiments show the requirement of a larger dataset for the Masked Language Modelling to boost f1 score. While the f1 score for the larger class improved, we do not see boost in f1 for the sexually explicit class and this is justified by the paucity of positive class data. Vocabulary extension though very promising also needs further research to handle better tokenization and embeddings.

References

[1] cross-encoder/ms-marco-minilm-l-4-v2. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-4-v2.

[2] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.

[3] roberta-base · hugging face. https://huggingface.co/roberta-base.

[4] xlm-roberta-large · hugging face. https://huggingface.co/xlm-roberta-large.

[5] unitary/unbiased-toxic-roberta · hugging face. https://huggingface.co/unitary/unbiased-toxic-roberta.

[6] unitary/multilingual-toxic-xlm-roberta · hugging face. https://huggingface.co/unitary/multilingual-toxic-xlm-roberta.

[7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746, 2019.

[8] Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview.

[9] Pan12 deception detection: Sexual predator identification. https://zenodo.org/record/3713280#.YloFWi-cZ70.

[10] Hate and abusive speech on twitter. https://github.com/ENCASEH2020/hatespeech-twitter.

[11] Offensive comments in brazilian web. https://github.com/leondz/hatespeechdata.

[12] Profanity keywords. https://github.com/rominf/profanity-filter/blob/master/profanity_filter/data/en_profane_words.txt.

[13] Detecting sexual harassment in text. https://github.com/yunsukim10/detecting-sexual-harassment/blob/master/Data/Harassment/sexual_harassment_data3.csv.

[14] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, November 2020. Association for Computational Linguistics.

A Appendix

A.1 Materials and Methods

Our pre-trained model basically has the same structure as BERT. Below we describe in detail the pre-training and fine-tuning process of our model.

A.1.1 Fine-tuning BERT

BERT-base model contains an encoder with 12 Transformer blocks, 12 self-attention heads, and the hidden size of 768. BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The sequence has one or two segments that the first token of the sequence is always [CLS] which contains the special classification embedding and another special token [SEP] is used for separating segments. For text classification tasks, BERT takes the final hidden state h of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of BERT to predict the probability of label c:

$$p(c|h) = softmax(Wh)$$

where W is the task-specific parameter matrix. We fine-tune all the parameters from BERT as well as W jointly by maximizing the log-probability of the correct label.

A.1.2 Vocabulary Extension

We noticed many keywords and key phrases related to sexually explicit were being tokenized very inefficiently. This resulted in innocuous utterances containing words like those keywords being predicted as sexually-explicit.

To help combat this, and in the same vein as (Beltagy 2019[14], Tai 2020[15]), we extended the original tokenizer vocabularies with these keywords to better align them with our task. Specifically, we included approximately 100 domain-specific keywords. We added three versions of each keyword: all lowercase, all capitalized, and with only the first letter capitalized. This resulted in approximately 300 new tokens and allowed us to account for different capitalizations observed in the data.

Though the tokenization step is trivial, care is needed at the model stage, since transformer-based architectures we have used, the first step for the model is essentially transforming each token to a high-dimensional embedding space. The pitfall with this approach being that the embeddings for the new words would be random by default, hence detrimental to model performance.

As a baseline, we tried simply extending the vocabulary of a RoBERTa model as described above and directly finetuning on the internal data. As expected, we observed a significant drop in performance compared to using the original vocabulary, with precision for the sexually-explicit class dropping from 0.93 to 0.88, and F1-score for the sexually-explicit class dropping from 0.93 to 0.85.

A.1.3 Optimizing New Tokens' Embeddings

For further optimizations, we borrowed strength from the embeddings in the original vocabulary to improve the new words' initialization. Specifically, for each added keyword, we tokenized it under the original vocabulary, and initialized its embedding as the mean of its resulting tokens. Even taking this initialization and directly fine-tuning led to a noticeable improvement, albeit still below baseline performance. This method resulted in precision and F1-score for the sexually-explicit class of 0.92 and 0.89, respectively.

Additionally, taking inspiration from (Jie 2020), we experimented with tokenizing each new word with a leading space included – e.g., for the new token "WWG1WGA" using " WWG1WGA" as input for the original tokenizer. We then average the embeddings of the original tokens, as above.

A.1.4 Pre-training RoBERTa

The RoBERTa model[3] is pre-trained in the general domain corpus. For a text classification task in a specific domain, such as sexually explicit classification, its data distribution may be different from BERT. Therefore, we can further pre-train RoBERTa with masked language model1 and next sentence prediction tasks on the domain-specific data. We stick to masked language modeling for

pre-training in our work as research has shown that NSP (next sentence prediction) is not always effective.

We perform In-domain pre-training, in which the pre-training data is obtained from the same domain of a target task. For example, there are several different sexually explicit datasets available, which have a similar data distribution. We further pre-train BERT on the combined training data from these tasks.

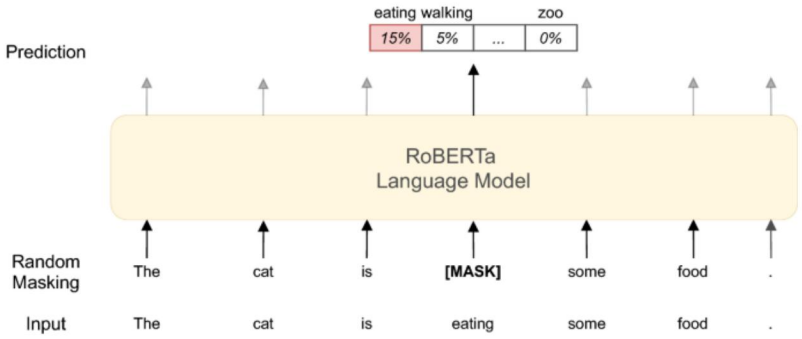

Figure 1: Masked Language Modeling.

A.1.5 Pre-training and Fine-tuning

Once we have a domain adapted language model via pre-training, we continue the fine-tuning task – sexually explicit classification, on the datasets selected (Table 1). Our work shows significant performance improvement when we continue the pre-training followed by the fine-tuning task, helping us reach state of the art in sexually explicit detection. Figure 2 shows three general ways for fine-tuning BERT

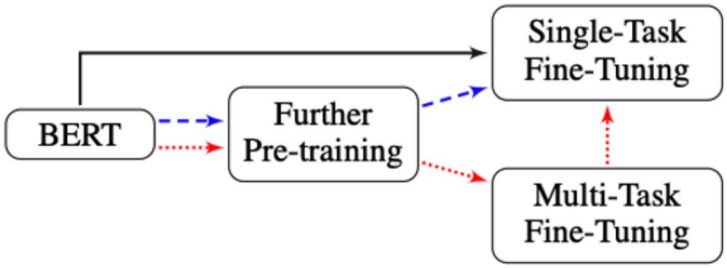

Figure 2: Three general ways for fine-tuning BERT, shown with different colors

A.1.6 Vocabulary Extension, Pre-training and Fine-tuning

We combined both our primary approaches - extended vocabulary and pre-training and fine-tuning the model. The model performance was below the best performing model from Masked Language Modeling. We see significant drop in model performance compared to pre-training and fine-tuning approach. We conclude that more research must be done in handling the tokens of the extended vocabulary and intelligent weighting of the embeddings.