Classification of Online Sexism With Different Transformer Models

Tony Lin
UCLA
CS 263
enlin@ucla.edu

Ryan Chien
UCLA
CS 263
ryanchien@ucla.edu

Peter Yang
UCLA
CS 263
yunping97@ucla.edu

Abstract

Sexism is a growing problem online that can cause harm and perpetuate injustices. While automated tools detect sexist content, they lack fine-grained classifications and explanations, reducing interpretability and trust. To empower users and moderators, we must identify and explain what constitutes sexist content. For this project, we will aim to develop more accurate and explainable English-language models for sexism detection, featuring 4 fine-grained classification models for Gab and Reddit's sexist content through 3 unique classification subtasks.

1. Introduction

This project was inspired by a CodaLab competition where contestants are given three unique classification subtasks with 20,000 social media comments from Gab and Reddit. The comments are to be classified as sexist or non-sexist, with some of the tasks being more granular than the others.

1.1. Motivation

Online sexism has always been a harmful tool used to inflict harm on women who are targeted. Classifying sexist comments has seen recent improvements with the development of automated tools. However, they are not accurate enough with classifications often being too vague. This can be especially frustrating for the victims whose livelihoods can sometimes depend heavily on the ability to detect these hateful comments. The goal of this project and paper is to build upon the previous shared tasks on abuse and hate detection and attempt to apply a taxonomy with three hierarchical tasks for detecting sexist content.

1.2. Background

As stated above, there are three hierarchical tasks for detecting sexist content for this classification problem. Task A is a binary task simply detecting if the content is sexist or not sexist. Task B distinguishes between four distinct categories of sexism: threats, derogation, animosity, and prejudiced discussion. Task C identifies one of 11 fine-grained sexism vectors. These vectors are better defined in Image 1.

2. Model Selection

Three models were used in this paper: BERT, GPT, RoBERTa, and XLNet. (Note: in the corresponding video submission we indicated we would use XGBoost as the third model. However, after further research we found RoBERTa and XLNet were better performing models and elected to use them instead).

2.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a natural language processing model initially introduced by Google. It implements transformer neural networks and is capable of understanding bidirectional context by utilizing self-attention, giving BERT the ability to better understand language nuances. BERT is normally pretrained and fine-tuned on a large corpus of unlabeled text using masked language modeling and next sentence prediction tasks. It generates contextual word embeddings, giving it the unique benefit of being able to provide contextually rich representations and assist in NLP downstream tasks.

In this project, we implement DistilBERT, a distilled version of BERT that is smaller, faster, cheaper and lighter. This allowed us to stay within the Google Collaboratory memory limit and still train the BERT model on the dataset. DistilBERT has the same general architecture as BERT, where the token-type embeddings and the pooling layer are removed and the number of layers is reduced by a factor of 2.

2.2. GPT

GPT (Generative Pre-trained Transformer) is a transformer neural network natural language processing model developed by OpenAI. Similarly to BERT, it utilizes self-attention and contextual word embeddings to generate human-like text. GPT is pretrained and fine-tuned on a large corpus of text where it predicts the next word in a sentence using a given preceding context. It excels in autoregressive generation; in other words, it predicts one word at a time based on previously generated words. This allows GPT to maintain coherence and provide contextually relevant sentences.

In this project, a standard GPT transformer model is used.

2.3. RoBERTa

RoBERTa (Robustly optimized BERT approach) is an improved version of BERT. It is trained with dynamic masking, where a new masking pattern is generated every time a sequence is fed into the model. It is also trained with full sentences without next sentence prediction loss, in large mini-batches, and a larger byte-level byte-pair encoding (a hybrid between character and word-level representation that allows handling of large vocabularies common in natural language).

2.4. XLNet

XLNet is a generalized autoregressive method that leverages autoregressive language modeling and autoencoding pre-training objectives. It addresses previous limitations of BERT by introducing a permutation-based training approach and applying a state-of-the-art Transformer-XL architecture. By maximizing the likelihood of predicting any word in a sequence, regardless of its position, XLNet captures dependencies among all words, overcoming the issue of context fragmentation. Context fragmentation refers to when a model lacks necessary context to predict the first few symbols in a sentence. Utilizing its autoregressive generation, transformer architecture, and contextual word embeddings, XLNet is better equipped to tackle various downstream NLP tasks.

3. Dataset

In this project, we are using the data provided by the CodaLab competition which consists of 20,000 social media comments from Reddit and Gab. These 20,000 samples are then distributed into three sets: training, development, test datasets for the models' performance purposes.

3.1. Data Collection

The current research on online harms heavily relies on Twitter data, which limits the diversity of sexist content in terms of its nature and severity. To overcome this limitation, the researchers opt to incorporate two platforms, Gab and Reddit, in their study. Gab is an alternative social networking site known for its emphasis on free speech, including far-right expressions. On the other hand, Reddit is a network of topic-based forums where users can engage with content aligned with their interests.

Equal amounts of data are collected from both Gab and Reddit, resulting in the creation of a pool containing 1 million entries for each platform. For Gab, a comprehensive dataset of 34 million publicly available Gab posts from August 2016 to October 2018 is collected. This dataset has been widely utilized in academic studies. From this dataset, a random sampling procedure is employed to select 1 million entries for the pool, which would then select 10,000 entries as our dataset.

Regarding Reddit, the researchers take a multi-step approach. They compile a list of 81 subreddits that are likely to contain sexist content based on previous research. Utilizing the Reddit API, all comments posted in these identified subreddits between August 2016 and October 2018 are collected. To ensure a well-balanced representation of linguistic expressions and topics, the focus is narrowed down to the 24 subreddits with at least 100,000 comments. Consequently, a dataset of 42 million comments is obtained. Finally, 250,000 comments are randomly sampled from each of the four subreddit categories, resulting in the creation of the final pool for analysis. From this pool, 10,000 entries are selected for labeling purposes.

3.2. Dataset Distribution

In this project, the dataset was partitioned into training, development, and test sets with the distribution ratio of 70%, 20%, and 10% respectively.

4. Experimental Setup

Based on the competition guidelines, we have separated our objectives into three tasks, and each of our models would perform and be evaluated on all three tasks. From simple to more complex, the tasks are binary classification, categorical classification, and the fine-grained vectors classification. The detailed definition of each tasks is shown in Image 1 below.

4.1. Task A – Binary Sexism

The first level of our objective is relatively simple. Based on the training data, we are trying to classify whether the test comment is or is not considered sexist.

4.2. Task B – Category of Sexism

The second level of the project task further breaks down sexist content into four distinct categories: threats, derogation, animosity, and prejudice discussion, which are all conceptually and analytically distinct.

4.3. Task C – Fine-Grained Vectors

Lastly, the last level which is also the most difficult one is the fine-grained vectors classification. Along with the category classified from task B, task C also disaggregates each category into fine-grained sexism vectors. There are 11 fine-grained vectors total, and each one of them are mutually exclusive to achieve optimal classification prediction (the 11 fine-grained vectors are detailed in Image 1).

5. Result

This section demonstrates the result of four different transformer models: GPT, BERT, RoBERTa, and XLNet. We have the accuracy, precision, recall, and F1 score for each model for performance comparison.

5.1. Performance of GPT

The accuracies, precisions, recalls, and F1 scores of GPT are shown in Table 1. For Task A, GPT can reach 86.33% accuracy, 85.79% precision, 86.33% recall, and 85.83% F1-score. On Task B, GPT comes 50.21% on accuracy, recall, and F1-score along with 51.04% precision. Lastly, GPT got 42.17% accuracy, 38.79% precision, 42.16% recall, and 39.70% F1 score on Task C.

5.2. Performance of BERT

The scores of BERT are shown in Table 1. BERT can get 86.03% accuracy, 85.71% precision, 86.03% recall, and 85.83% F1 score on Task A. For Task B, it reaches 56.60% accuracy and recall, 56.36% on precision, and 56.47% on F1 score. Finally, 47.11% accuracy, 43.58% precision, 47.11% recall, and 44.93% F1 score are the scores for BERT on Task C.

5.3. Performance of RoBERTa

RoBERTa can get 86.03% accuracy, 85.71% precision, 86.03% recall, and 85.83% F1 score on Task A. For Task B, it gets 58.76%, 58.15%, 58.76%, and 57.72% on accuracy, precision, recall, and F1 score, respectively. Last, on Task C, it reaches 51.24% accuracy, 49.62% precision, 51.24% recall, and 49.35% F1 score.

5.4. Performance of XLNet

For Task A, XLNet can get 87.23% accuracy, 86.98% precision, 87.23% recall, and 87.07% F1 score. For Task B, 58.14% accuracy, 57.83% precision, 58.14% recall, and 57.35% F1 score are reached. Finally, for Task C, it got 50.93% accuracy, 50.01% precision, 50.93% recall, and 49.94% F1 score. A thing we noticed is that in terms of training time, XLNet takes a noticeably longer time to finish the training process.

6. Discussion

In this section, we will discuss the model performances and comparisons according to the differences between these four models. Notice that the comparisons are based on online sexism classification but not other tasks. Moreover, due to resource limitation, we use XLNet, which is the only non-distilled model we can run without having memory issues, as the relatively ideal model for comparison.

6.1. Performance On Easier Tasks

As shown in Table 1, on Task A, a relatively easy task, these three distilled models perform similarly. Although the performance of GPT is the best among the other distilled models, the difference is not large. We believe the ranking between these three distilled models on this task can vary in different trials. That is, for simple tasks, different transformer models do not have significant differences. However, XLNet outperforms the other models, showing the gap between a distilled and non-distilled version of the model on easier tasks.

6.2. Performance On Harder Tasks

On Task B and C, which are relatively harder tasks, the difference between different models starts to become larger. As RoBERTa is the better variant of BERT, it is obvious that the former performs significantly better than the latter, with a 1.25% difference in F1 score on Task B and a 4.42% difference in F1 score on Task C, not to mention that BERT has already outperformed GPT, with a 6.26% difference in F1 score on Task B and a 4.95% different in F1 score on Task C. We will discuss the explainable reasons why RoBERTa outperforms the other distilled models in another section. More importantly, we notice that XLNet does not outperform the other distilled models on harder tasks.

6.3. Number of Parameters

In our experiment, due to resource limitations, we have to use the distilled version of GPT, BERT, and RoBERTa. According to Hugging Face, DistilGPT2 and DistilRoBERTa have 82 million parameters, and DistilBERT has 66 million parameters. The only complete model that we can run is XLNet, which has 110 million parameters. With the information above and the result of the experiment, it shows that the number of parameters is not the crucial reason why a model can have better performance. However, it is the key differences they have in their architecture and training objectives that contribute to their varying performance on online sexism classification.

6.4. Model Architecture

Although these three distilled models are all based on the transformer architecture, they have different modifications. GPT uses a decoder-only architecture, where the self-attention mechanism is used to generate the next word in the sequence. In contrast, BERT and RoBERTa use a bidirectional transformer, which allows them to consider both left and right context during training. This helps capture a richer representation of the text, resulting in better performance for these two models.

Moreover, as RoBERTa is a variation of BERT that has been further optimized, it was trained on a much larger corpus of data compared to BERT. In addition, it also uses dynamic masking, which randomly samples different masks for each epoch, different from BERT using static masks during pretraining. Furthermore, RoBERTa simplifies the training objective by removing the next sentence prediction tasks. These modifications are simple but very effective, letting RoBERTa outperform BERT as the result shows.

In addition, although XLNet is the only non-distilled model, the performance is not superior to the distilled version of RoBERTa. The other downside of XLNet is that it takes a longer time to train. Compared to DistilRoBERTa's performance and speed, it does not seem like it is worth training with such a huge model with only a little better performance on easy tasks.

Overall, these four models have their strengths and perform well on various tasks. However, for online sexism classification, when considering the balance between performance and computation efficiency, RoBERTa leads itself to perform better than the other models.

Acknowledgment

The content of this paper is used in CS 260C for a similar subject matter.

References

[1] Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023, May 8). Semeval-2023 task 10: Explainable detection of online sexism. arXiv.org. https://arxiv.org/abs/2303.04222

[2] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[3] Li, Tianda, et al. "A short study on compressing decoder-based language models." arXiv preprint arXiv:2110.08460 (2021).

A. Performance of Different Models

		GPT	BERT	RoBERTa	XLNet
Task A	Acc	86.33%	86.03%	85.63%	**87.23%**
	Pre	85.79%	85.71%	85.87%	**86.98%**
	Rec	86.33%	86.03%	85.63%	**87.23%**
	F1	85.83%	85.83%	85.73%	**87.07%**
Task B	Acc	50.21%	56.60%	**58.76%**	58.14%
	Pre	51.04%	56.36%	**58.15%**	57.83%
	Rec	50.21%	56.60%	**58.76%**	58.14%
	F1	50.21%	56.47%	**57.72%**	57.35%
Task C	Acc	42.17%	47.11%	**51.24%**	50.93%
	Pre	38.79%	43.58%	49.62%	**50.01%**
	Rec	42.16%	47.11%	**51.24%**	50.93%
	F1	39.70%	44.93%	49.35%	**49.94%**

Table 1: Accuracies, Precisions, Recalls, and F1-Scores of Different Models

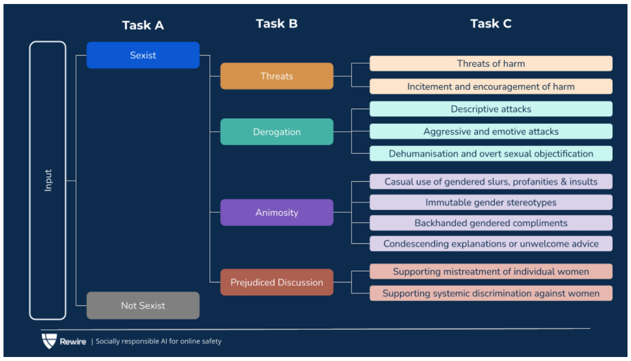
Task A
Task B
Task C
Input
Sexist
Not Sexist
Threats
Derogation
Animosity
Prejudiced Discussion
Threats of harm
Incitement and encouragement of harm
Descriptive attacks
Aggressive and emotive attacks
Dehumanisation and overt sexual objectification
Casual use of gendered slurs, profanities & insults
Immutable gender stereotypes
Backhanded gendered compliments
Condescending explanations or unwelcome advice
Supporting mistreatment of individual women
Supporting systemic discrimination against women
Rewire | Socially responsible AI for online safety

Image 1: Project objectives with three levels of tasks