An Analysis of Fine-Tuning RoBERTa for Content Moderation on Reddit

Ray Mohabir
Center for Data Science
New York University
rm3887@nyu.edu

Anusha R. Patil
Center for Data Science
New York University
arp624@nyu.edu

Zane David Dennis
Center for Data Science
New York University
zdd210@nyu.edu

Abstract

A significant problem since the dawn of social media has been content moderation–most notably the removal of threatening, violent, harassing, or otherwise inappropriate content posted by users. Historically, this has typically required large teams of human moderators relying on user "reports" that flag content for manual review. However, these methods are both work-intensive and low-recall. This problem is exacerbated on a social network like Reddit, where each "subreddit" community has a different set of rules and standards regarding what is acceptable content. So even a hypothetical automatic moderation tool that performs perfectly on one subreddit might be unhelpful on another. Deep learning, such as CNNs and RNNs, has shown promise in automatic comment flagging, and we plan to continue this work using newer models such as RoBERTa. We propose a RoBERTa model fine-tuned on a labeled dataset of flagged comments from a specific subreddit to create an automatic comment-flagging model specific to the rules of a given subreddit.

1 Introduction

Social media has become a crucial platform in modern society for everything from sports to politics to entertainment. For example, Jack Dorsey, founder of Twitter, was quoted in (Rogers, 2013) describing his site as having changed from "a what-I-had-for-lunch medium" to "an event-following tool." In 2018, a Pew Research Poll Shearer (2018) found that social media now "surpasses print newspapers as a news source for Americans." However, one rampant problem in the new digital frontier is content moderation: the decisions of what content will be allowed on the site. As described by Cheng et al. (2015), many users often use anonymous social media platforms to harass, bully, or exhibit otherwise antisocial behavior. Such behavior can even include "illegal content, copyright violations, terrorism and extremist content, revenge porn, online harassment, hate speech, and disinformation" (Jhaver et al., 2019). As documented by Mahajan et al. (2020), unchecked "cyberbullying" can have drastic consequences on its victims, even culminating in suicide. While content moderation is a three-pronged issue–ethical, legal, and technical–in this paper we focus on the technical challenges of enforcing established rules on a massive social media platform.

One such social media site is Reddit, which consists of topic-focused communities known as "subreddits" ranging from /r/politics to /r/gaming to /r/nba (American professional basketball). While there are a small set of site-wide rules known as "Reddiquette" that essentially boil down to "be nice and don't do anything too illegal," declaration and enforcement of community-specific rules is left to the discretion of volunteer moderators appointed by the founder of a subreddit or previous moderators. In this experiment, we focus on /r/nba (`https://www.reddit.com/r/nba`), as it is one of the largest communities on Reddit (over 3 million subscribers) and also poses the unique challenge of community- and domain-specific jargon. Not only do comments often reference previous incidents ("3-1" is a common reference to the Golden State Warriors' embarrassing Finals loss in 2016, and is often used to maliciously troll their fans to this day), they also include slang like "nephew," a common insult generally declaring the recipient to be immature, young, and/or stupid (`https://knowyourmeme.com/memes/delete-this-nephew`).

Content moderation on Reddit and many similar sites is a two-step process consisting of a "report" stage and a "modqueue" stage. The report stage consists of users flagging any content on the site they believe is against the rules; this is in essence a high-recall/low-precision binary classifier. These reports are then sent to the "modqueue," where human moderators make final decisions on whether or not the content should be removed from the site (https://mods.reddithelp.com/hc/en-us), acting as a higher precision binary classifier. While the user reports are highly inaccurate, as users are not as familiar with the rules and can (ironically) report any content they don't like for petty or trolling reasons, the user report stage is necessary to process the massive throughput of activity on these communities. Our case study of /r/nba, for example, has a team of more than 30 human moderators. Even so, user reports often fail to capture large swaths of rule-breaking content, which then build to higher, more disruptive levels before they are finally flagged.

The primary built-in automation tool that Reddit gives its moderators is known as Automoderator (Automoderator, 2018), which at its most complex allows custom regex queries (written by the moderators) to flag unwanted content. Automoderator acts primarily at the report stage but can also be configured to automatically remove the content in question, skipping the modqueue stage entirely. The digital nature of these discussion boards makes their moderation an excellent candidate for AI automation or supplementation, as modern machine learning and deep learning techniques are far more sophisticated than Automoderator's existing toolbox. The diverse nature of Reddit's various subreddits also provides a challenging proving ground for generalizable automated moderation techniques. In this experiment, we propose a high-recall classifier capable of supplementing Automoderator and user reports to identify problematic content instantly as it is posted.

Our main contributions are as follows: 1) according to our research, this study is the first to investigate fine-tuning RoBERTa for content moderation on social media, and 2) our experimental model achieves a recall of 100% at a precision of 74% or a recall of 88% at a precision of 83%, depending on the configuration. We conclude by suggesting future work to enhance the performance and usefulness of our model.

2 Related Work

Pavlopoulos et al. (2017) show the effects of using word embeddings on CNNs and N-gram models, finding that N-gram models perform considerably worse than CNN models. In addition, they discuss the effects of different input word representations such as GloVe embeddings and word2vec for various NLU tasks. In light of this, we used a BiLSTM as our baseline so as to apply GloVe embeddings to our work on content moderation.

Mahajan et al. (2020) talk about the effects and approaches used to tackle increasingly common abusive and inappropriate comments on social media. In order to classify whether a comment was abusive, they ran Logistic Regression, Multinomial Naive Bayes, Random Forest, Xg-boost, CNN, and GRU models using preprocessing techniques such as TF-IDF and GloVe Embeddings. They explain that using GloVe Embeddings gave the most intuitive explanations for the predictions being made. This again led us to use word embeddings such as GloVe and Transformers embeddings in our models.

Dadvar et al. (2013) describe the importance of including user contexts such as post history, profile information, and user characteristics that traditional content moderation disregards. While these features are important in content moderation, social media, and in particular Reddit, typically have user-driven self-moderation in the forms of likes/dislikes and upvotes/downvotes that allow machine moderators, like AutoModerator from Reddit, to ban the user's comments. While this disregards the sentiment behind individual posts, being banned by AutoMod for having too many negative comments gives a rough indication of the user's post history. As such, these features were not included in our dataset for now because they are addressed through other means. However, future work might pursue doing so.

Davidson et al. (2017) explain the challenge of determining the intent behind offensive language and separating it against hate speech in general. Davidson et al crowdsourced their dataset and produced a hate speech lexicon of tweets that were labeled as either hate speech, offensive language, or neither of them. Using Logistic Regression, they found that their model misclassifies hate speech 40% of the time and suggests that the Logistic Regression model was biased towards offensive language. This challenge will be heavily prevalent in our experiment, as two users discussing an incident in which one NBA player calls another a slur is vastly different than a user calling another user that same slur.

Xu et al. (2019) propose an improved word representation method in a BiLSTM model to capture sentiment context behind comments. In addition, Xu et al compared their sentiment analysis method to other methods including RNN, CNN, LSTM, and NB. They determined BiLSTM fully understands underlying context information in comments and can better represent the sentiment in user comments. From this, we reinforced our decision to use BiLSTM as a baseline model so that we can compare our experimental model to one with understanding of context.

3 Data

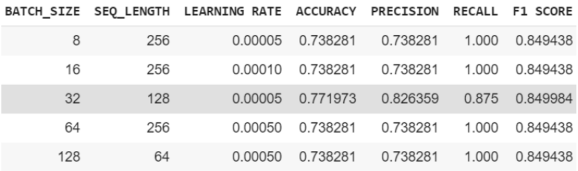
Our data was scraped from the private moderator logs of /r/nba, filtering specifically for comments that were either removed or approved (as opposed to posts or other moderator actions). All comments were flagged for manual review by either a user or Automoderator. The dataset consists of the full text of 20,000 user comments (10,000 approvals and 10,000 removals), with the binary labels “approvecomment” and “removecomment” which were translated into 0 and 1, respectively. The data along with baseline implementations can be found in the public GitHub repository. [1]

4 Methods

BiLSTM: We first implemented a baseline model using BiLSTM. A BiLSTM uses two LSTMs to learn each token of the sequence based on both the past and the future context of the token. One LSTM processes the sequence from left to right, the other one from right to left. The forward and backward context representations are concatenated into a long vector. Using GloVe embeddings, we created 300-dimensional word vectors, where only the top 50000 word embeddings were used from GloVe. We chose 300-dimensional embeddings because the performance is good without taking too much time to train. After minimal data preprocessing, we proceeded to the next step of tokenization using Moses Tokenizer from Sacremoses. For the BiLSTM, we loaded the weights from the GloVe embeddings and set the maximum sentence length to a static value of 256. After performing hyperparameter tuning on the LSTM classifier, we get an accuracy of 0.7466 on our validation set. The model is able to classify the Reddit comments into approved and removed comments reasonably well with minimal preprocessing of the text data.

RoBERTa: For our experimental model, we use RoBERTa-base (Liu et al., 2019) on the labeled dataset. Using the embeddings from a pretrained model from Transformers (Wolf et al., 2020), we achieve a result showing how well RoBERTa-base deals with identifying underlying context behind user comments and correctly identifying whether the comment should be removed. To ensure the dataset was able to be evaluated, the data was minimally preprocessed by removing all tab delimiters in each comment, allowing the program to read in and distinguish user comments and labels. Since RoBERTa-base was used, the labelled dataset was tokenized using Transformers RobertTokenizer method. This method uses a byte-level byte-pair encoding to encode Reddit comments into compressed subwords. The experiment was done was on the following hyperparameter values and was ran on five epochs: 1) batch sizes of [8, 16, 32, 64, 128], 2) max sequence lengths of [32, 64, 128, 256], and 3) learning rates of [1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3].

5 Results & Analysis



BATCH_SIZE	SEQ_LENGTH	LEARNING RATE	ACCURACY	PRECISION	RECALL	F1 SCORE
8	256	0.00005	0.738281	0.738281	1.000	0.849438
16	256	0.00010	0.738281	0.738281	1.000	0.849438
32	128	0.00005	0.771973	0.826359	0.875	0.849984
64	256	0.00050	0.738281	0.738281	1.000	0.849438
128	64	0.00050	0.738281	0.738281	1.000	0.849438

Figure 1: Best F1 results for each batch size

As shown in Fig 1, our model achieves very high recall values, even reaching 1.0 in some configurations, while still reaching somewhat high precision values of around 0.74 (or up to 0.83 in configurations with lower recall). While the relatively lower precision values mean that the model is not yet intelligent enough to replace human moderating at the ”modqueue” stage entirely, the high recall values indicate that our model would perform extremely well at the ”report” stage of the process. This would lead to more thoroughly enforced rules, as moderators would have a much greater portion of rule-breaking content brought to their attention. It would also lead to quicker removal of illicit content,

[1] https://github.com/anushapatil5/NLU-ContentModeration

as the model could process all comments immediately upon posting instead of waiting for a user to see it and decide to report.

It is notable that several of our results are identical. We believe one cause of this is that batch size and sequence length are largely inconsequential hyperparameters to this problem, as removed comments are typically quite short. Tiny learning rate values would then enable each model setup to reach the same local minimum, at a granularity smaller than that necessary to produce different output in the validation set (which was rather small, at 2048 samples).

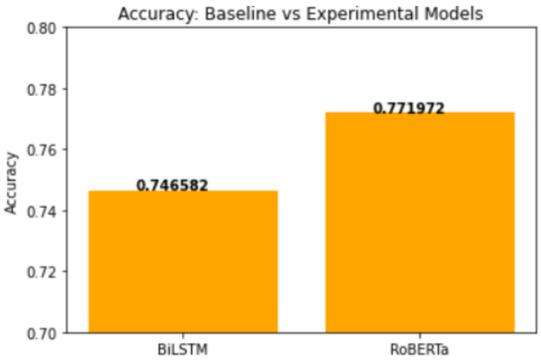


Figure 2: Comparison of best accuracy between baseline and experimental models

Interestingly, our experimental RoBERTa model does not show a very large improvement in accuracy over the baseline BiLSTM model, even though RoBERTa is generally a much better performing model on NLU tasks. We see two likely explanations for this. First, that our performance may be approaching a limit of human "accuracy" in labeling. Since our dataset came from real-world moderator actions, it is likely that some moderators are stricter than others even when enforcing the same rules, thus leading to noisy labeling. Emotional responses by human moderators could also contribute to inconsistent decisions, as discussed in (Jhaver et al., 2019). Second, that our work has not taken advantage of the full capabilities of RoBERTa such as pretraining with unlabeled data to gain a more accurate understanding of the platform's domain-specific jargon, as was done in SciBERT (Beltagy et al., 2019).

6 Conclusion & Future Work

In this paper, we present an analysis on fine-tuning RoBERTa with labeled comments from a specific subreddit to create an automatic comment-flagging model specific to the rules of that subreddit. While our model is not yet capable of replacing human moderators, its high recall values show that it is already capable of being an excellent flagging system. We suggest that future work should focus on increasing the precision of the model so as to create a fully self-contained moderation system capable of synthesizing both stages into one. Our first suggestion, to address the concern of inconsistent labeling, is to have multiple human moderators vote on each label so as to reduce noise around what is considered an approved comment or a rule-breaking comment. Our second is to pursue pretraining of RoBERTa as described above; this improvement would be particularly useful on a platform like Reddit where each "subreddit" community has its own unique jargon and rules; a model for use on a specific subreddit could be pretrained with text from that subreddit (in addition to the fine-tuning on comment removal) to create an even more intelligent classifier for that subreddit's content. We also suggest that a similar subcommunity-centric approach to natural language processing should be explored on other social media platforms such as twitter, where users do not have official communities like subreddits but nevertheless cluster unofficially into overlapping communities with their own jargons and histories.

7 Collaboration Statement

Zane acquired the data, performed data cleaning and literature review, and contributed to experimental model bugfixing and results analysis. Anusha performed data preprocessing, implementation of baseline model (BiLSTM) and LaTeX documentation. Ray conducted literature review, related work summaries, and fine-tuning of RoBERTa. All members contributed equally to preparing the presentation, full paper, and overall discussion.

References

Automoderator. 2018. Automoderator - reddit.com. retrieved from `https://www.reddit.com/wiki/automoderator`.

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. ArXiv:1903.10676.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. *Proceedings of the Ninth*

International AAAI Conference on Web and Social Media.

Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. “improving cyberbullying detection with user context. *Advances in Information Retrieval*.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. ArXiv:1703.04009.

Jhaver, Shagun, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *Association for Computing Machinery*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv:1907.11692.

Mahajan, Aditya, Divyank Shah, and Gibraan Jafar. 2020. Explainable ai approach towards toxic comment classification. *EasyChair*.

Pavlopoulos, John, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. *Association for Computational Linguistics*.

Richard Rogers. 2013. Debanalizing twitter: the transformation of an object of study. *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference*.

Elizabeth Shearer. 2018. Social media outpaces print newspapers in the u.s. as a news source. *Pew Research Center*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, and Pierric Cistac. 2020. Huggingface's transformers: State-of-the-art natural language processing. ArXiv:1910.03771.

Xu, Guixian, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. 2019. Sentiment analysis of comment texts based on bilstm. *IEEE Access 7*.