AI for Language: Community Rules Classification

Pan, Huangyu
1155211012@link.cuhk.edu.hk

Weng, Xiangxiang
1155211173@link.cuhk.edu.hk

Abstract

This report details a progressive natural language processing project aimed at developing a robust binary classifier to predict rule violations in Reddit comments. Facing challenges such as heterogeneous moderation standards and data scarcity regarding "unseen" rules, we implemented and evaluated three distinct methodological approaches: a foundational TF-IDF with Logistic Regression baseline, a high-performance Large Language Model (LLM) fine-tuning solution, and a Zero-Training Semantic Search method. The key to achieving state-of-the-art generalization was the advanced LLM solution, which employed efficient QLoRA fine-tuning (using Qwen3-14B), Logits Pooling to convert the generative model into a precise classifier, and a Multi-LLM Ensemble. The final ensemble model achieved a competitive Column-Averaged AUC of ≈ 0.93, successfully demonstrating strong generalization capabilities necessary for real-world moderation systems.

1 Introduction

Online community platforms, such as Reddit, rely heavily on moderation systems to enforce rules and maintain community standards. However, the decentralized nature of these platforms, where each subreddit operates under its own jurisdiction, often results in inconsistent and unpredictable enforcement. This project [1] addresses the challenge of predictive content moderation by developing a machine learning model capable of estimating the likelihood that a given comment, paired with a specific rule, constitutes a violation.

row_id	body	rule	positive_example_1	positive_example_2	negative_example_1	negative_example_2	rule_violation
0	Banks do	No Adver	If you could tell	hunt for lady for ja	Watch Golden Globe	DOUBLE CEE x BANDS	0
1	SD Strea	No Adver	[I wanna kiss you al	LOLGA.COM is One of	#Rapper	[15 Amazing Hidden F	0
2	Lol. Try	No legal	Don't break up with h	It'll be dismissed:	Where is there a si	Because this stateme	1
3	she will	No Adver	Selling Tyrande	tight pussy watch fc	NSFW(obviously) htt	Good News ::Downloac	1
4	code	No Adver	wow!! amazing remin	seek for lady for se	must be watch movie	We're streaming Poke	1
5	I live i	No legal	Public School? You a	OP, you need to get	It looks like it co	That is called batte	0

The input data for this binary classification task consisted of a Reddit comment body, the target rule, and a crucial few-shot context comprising two positive (violating) and two negative (non-violating) examples. The evaluation metric used throughout the project and for final scoring was the Column-Averaged Area Under the Curve (AUC). To assess generalization power, the test set was strategically divided into a Public Leaderboard (30%) for immediate feedback and a Private Leaderboard (70%) for final, unbiased evaluation.

2 Methodology and Model Progression

To address the dual challenges of data scarcity and rule semantic complexity in community rule text classification, this study follows a principle of progression from simplicity to sophistication. We designed and evaluated three representative technical approaches. Systematically moving from

traditional feature engineering to advanced solutions based on Large Language Models (LLMs), we explored the differences in semantic understanding depth and few-shot generalization capability among these methods.

2.1 Model I: Initial Baseline (TF-IDF and Logistic Regression)

As an exploratory baseline, we first implemented a classic Natural Language Processing (NLP) pipeline. Textual features were encoded using the Term Frequency-Inverse Document Frequency (TF-IDF) scheme. This method quantifies word importance by balancing its frequency within a document (Term Frequency) against its rarity across the entire corpus (Inverse Document Frequency).

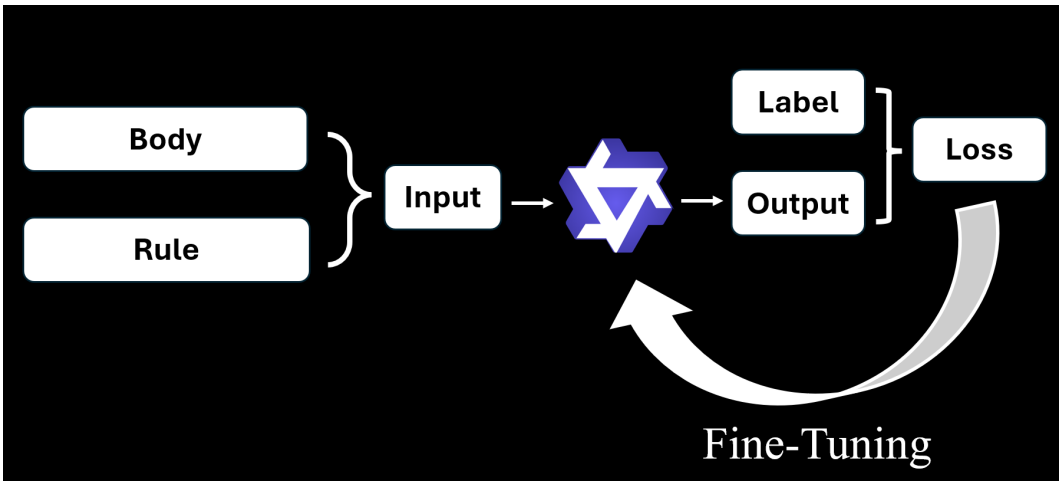
The resulting TF-IDF feature vectors were then fed into a Logistic Regression classifier for training and prediction. This model established a clear performance benchmark but was fundamentally limited: TF-IDF is essentially a variant of the Bag-of-Words model, treating words as independent features. This approach fails to capture word order, dependencies, and deeper contextual semantics, which are critical for judging complex and nuanced rule violations. The limitations of this model clearly indicated the necessity of transitioning to deep learning methods capable of semantic understanding. The model achieved a Column-Averaged AUC of approximately 0.54, confirming its utility as a baseline while highlighting the performance ceiling of non-contextual, statistical feature representations.

2.2 Model II: Advanced LLM Fine-Tuning (Top 1 Solution)

Recognizing the superior performance of Large Language Models (LLMs) in this domain, our primary focus shifted to implementing and optimizing a state-of-the-art fine-tuning approach, based on the top-performing solutions in the competition [2].

2.2.1 Data Preparation and Few-Shot Training

The core challenge of the competition was generalizing to rules not seen in the initial training set. The ideal scenario involves In-Context Learning (ICL), where the model sees the few-shot examples at inference time. However, to maximize efficiency and generalize robustly, we adopted an advanced strategy to transfer ICL knowledge into the model's weights during fine-tuning.



Instead of encoding all columns (`body`, `rule`, `positive_example_1/2`, `negative_example_1/2`) into every single prompt, the final training and inference prompt structure was kept minimalist, encoding only the `body` and the `rule`. Crucially, the few-shot examples were used to create synthetic training samples. The texts from the `positive_example_1/2` and `negative_example_1/2` columns in `test.csv` were extracted, labelled with 1 or 0, and up-sampled to create new, independent training rows. The example text was mapped to the `body` column in these new rows. This technique allowed the model to learn the semantic boundaries of the new rules by training heavily on their specific examples, thereby encoding the ICL capability directly into the Qwen3-14B weights without relying on a large, complex prompt at runtime.

2.2.2 Loss Masking Strategy

To ensure the model learned the specific classification task rather than simply memorizing prompt patterns, we utilized Loss Masking during fine-tuning. By applying the `train_on_responses_only` utility from Unsloth [3], the Cross-Entropy Loss was only calculated over the 'Yes' or 'No' answer tokens generated by the model, while the loss from the input prompt (including the rules and examples) was masked (set to zero). This forces the model to encode the semantic boundary of the rule into its weights.

2.2.3 Logits Pooling and Classification

Since the LLM (Qwen3-14B) is inherently a generative model, we introduced the Logits Pooling method to convert its output into a stable binary probability without generating extraneous text. First, We halt the model's forward pass immediately after the final Transformer layer, before the first token is generated, extracting the raw Logits vector corresponding to the entire vocabulary. Then, we identify the vocabulary IDs for all positive variants (e.g., 'Yes', 'True') and negative variants (e.g., 'No', 'False'). Finally, We normalized them using Softmax, to produce the final probability of a violation.

This approach improves inference efficiency by relying on a single `forward()` pass.

2.2.4 Multi-LLM Ensemble

The final prediction was derived from a weighted average ensemble of six separate fine-tuned models. This ensemble strategy [4] consistently outperformed any single model, delivering a robust final score.

2.3 Model III: Zero-Training Semantic Search

As an alternative to the fine-tuning approach, we explored a Zero-Training Semantic Search method [5]. This method relies entirely on the quality of a pre-trained embedding model and requires no task-specific training. In this scheme, all texts (training and test data, including rules and examples) are converted into dense vectors via the pre-trained encoder. For any new query (comment and rule), the system retrieves the K nearest neighbor embeddings based on vector similarity. The final violation score is computed by aggregating the similarity scores of these nearest neighbors, weighted by their labels. In our implementation, we used the Qwen 3-0.6B-Embedding model for generating embeddings and set K=1000 for neighbor retrieval. This method effectively addresses the semantic limitations of the baseline model while avoiding the computational cost associated with fine-tuning. This approach achieved a significantly higher AUC of approximately 0.87. Notably, its performance on the Public and Private Leaderboards was remarkably close, indicating a well-generalized solution with minimal overfitting to the public test subset.

3 Conclusion

This project successfully implemented and evaluated a series of models for predictive content moderation, highlighting the profound performance disparity between traditional methods and fine-tuned LLMs. The progression from a semantically-limited TF-IDF baseline to a sophisticated Qwen3-14B model demonstrated that high performance is achieved through meticulous technical application. Specifically, the adoption of Few-Shot fine-tuning, Logits Pooling, and Loss Masking were critical innovations that transformed a generative model into a robust, high-precision classifier capable of generalizing across diverse and previously unseen community rules. This work underscores that deep understanding of LLM fine-tuning mechanisms is essential to achieve robust, competitive results in complex real-world NLP tasks.

References

[1] J. Sorensen, L. Dos Santos, L. Vasserman, M. Cruz, T. Acosta, and W. Reade. Jigsaw - Agile Community Rules Classification. Kaggle. 2025. https://kaggle.com/competitions/jigsaw-agile-community-rules.

[2] G. Xu. 1st place solution. Jigsaw - Agile Community Rules Classification Solution Writeup, Kaggle. 2025. https://www.kaggle.com/competitions/jigsaw-agile-community-rules/writeups/1st-place-solution.

[3] K. Singhapoo, A. Inthanil, and A. Pillai. Fine-Tuning AI Models with Limited Resources. In 2025 11th International Conference on Engineering, Applied Sciences, and Technology (ICEAST), pages 148–151, 2025.

[4] Z. Chen, J. Li, P. Chen, Z. Li, K. Sun, Y. Luo, Q. Mao, M. Li, L. Xiao, D. Yang, Y. Ban, H. Sun, and P. S. Yu. Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. arXiv preprint arXiv:2502.18036, 2025.

[5] K. Tushin. Just use semantic search. Qwen3-emb-0.6B. Kaggle Notebook. 2025. https://www.kaggle.com/code/neibyr/30-min-just-use-semantic-search-qwen3-emb-0-6b/notebook.