Guarding Consumers with NLP: Extracting Risk Signals from CFPB Complaint Narratives

Alex Kim & Patrick Abousleiman
School of Information, University of California, Berkeley

Abstract

The CFPB publishes large volumes of consumer complaint narratives that historically contributed to supervisory oversight, but recent reductions in regulatory activity leave these signals underutilized. This work evaluates whether current NLP models can identify complaints indicative of consumer harm or regulatory violations, using company relief outcomes as a weak supervision signal. We compare a TF–IDF logistic regression baseline with several transformer architectures, including LoRA-FinBERT, fully fine-tuned FinBERT, and RoBERTa/ModernBERT variants—under severe class imbalance (risky cases ~5%). While the baseline yields a strong ROC–AUC, full FinBERT and RoBERTa yield higher minority-class precision and more stable decision boundaries. Post-hoc error and clustering analyses reveal that models detect harm-oriented narratives but struggle with rule-based violations expressed in procedural language. These findings underscore both the potential and the limits of narrative-only approaches and motivate retrieval-augmented methods that inject regulatory knowledge.

1 Introduction

The Consumer Financial Protection Bureau (CFPB) maintains one of the largest public repositories of financial complaints in the United States. These free-text narratives provide timely signals about emerging risks in consumer finance. However, with the CFPB reducing its supervisory activity in 2025, millions of incoming complaints now lack a corresponding mechanism for systematic review (Krause, 2025). This shift underscores the need for automated methods to identify narratives indicative of consumer harm or regulatory violations.

This paper investigates whether modern NLP models can support such early-warning detection. We frame the task as binary classification using outcome labels—'closed with monetary relief' or 'closed with non-monetary relief'—as weak but meaningful proxies for regulatory risk. A central challenge is extreme class imbalance: only about 5% of complaints receive any relief. Prior work has shown that such an imbalance makes accuracy misleading and suppresses minority-class learning (Roumeliotis et al., 2025; Oyewola et al., 2023).

The existing literature on CFPB data spans topic modeling (Bastani et al., 2019), deep learning classifiers (Oyewola et al., 2023), transformer-based topic extraction (Vasudeva Raju et al., 2022), and recent zero-shot reasoning approaches (Roumeliotis et al., 2025). Yet none address the specific task of predicting relief-based risk labels. We extend this line of work by evaluating several supervised NLP architectures, including a TF–IDF logistic regression baseline, LoRA-FinBERT, full FinBERT fine-tuning, and RoBERTa/ModernBERT variants, to determine whether transformer models meaningfully improve risk detection in long, domain-specific complaint narratives.

Our findings show that although the baseline classifier achieves a competitive ROC–AUC, transformer models substantially improve calibration and minority-class precision. Full FinBERT and RoBERTa in particular exhibit more stable performance under class imbalance. Through error and topic-cluster analyses, we also uncover systematic failure modes. Models reliably detect harm-oriented narratives but under-detect rule-based violations expressed in neutral language. These insights highlight both the promise and the limitations of supervised NLP for consumer-risk monitoring in the absence of active regulatory oversight.

2 Background

Much of this study builds upon previous work done relating to complaint analysis of CFPB data. Substantial research has been conducted involving different machine learning techniques to help identify how consumer complaint data analysis can be automated in a way that is both accurate and cost-effective for the agency. This task has become especially critical in recent months, given the current downsizing of the agency and expected increase in consumer complaint volume.

Initial research into this topic was conducted in 2019 using Latent Dirichlet Allocation (LDA), a modeling approach used to identify hidden patterns within a dataset. LDA analyzes the text of several narratives to identify common words in the data known as “topics”, and can then assign several topics to each complaint in order to circumvent the single-issue limitation of the CFPB’s original approach. The study found 40 distinct topics in the dataset, such as “Credit Reporting” or "Harassment", that it was then able to use in labeling examples. By tracking how prevalent these topics are over time, the model could evaluate the effectiveness of the CFPB in regulating and enforcing their policies regarding these categories. This method also surfaced the 2015 RushCard outage, where thousands of customers temporarily lost access to funds (Bastani et al., 2019).

Subsequent research into classification techniques leveraged modern machine learning methods such as deep learning and transformer-based models to improve overall accuracy. One 2023 study took advantage of supervised classification using deep learning algorithms. The researchers used an innovative two-stage residual one-dimensional convolutional neural network (TSR1DCNN) to classify complaints into 11 predefined categories. This model yielded a test accuracy of 76.53%, outperforming other deep learning architectures (Oyewola et al., 2023). Another study sought to address weaknesses in the initial LDA approach to classification by attempting to capture deeper semantic relationships between words in the dataset. This study compared the original LSA and LDA models to more modern transformer-based models, including BERTopic, DistilBERT, RoBERTa, and FinBERT, a model pretrained on financial text. Of these models, FinBERT achieved the highest performance with a C_V score of 0.3327, while BERTopic models generated more meaningful, diverse, and easily interpretable topics. These improvements demonstrate the significant advantages of using domain-specific pretrained models for classification (Vasudeva et al., 2022).

The most recent studies conducted in this topic show how researchers explored ways in which the most innovative LLMs could use zero-shot classification to classify complaints into five categories without prior task-specific training. This experiment used novel specialized reasoning models capable of performing complex, multi-step inference. The study tested and evaluated how 14 different models were able to perform on the dataset and highlighted the tradeoffs made in accuracy, cost, and speed. The results indicate that the o1 reasoning model yields the highest accuracy at 0.777, but also has by far the highest cost of any of the models tested, while also being relatively slow. This makes it impractical for this use case compared to a model like Gemini-2.0-Flash, which recorded a slightly lower accuracy of 0.758, but at a much lower cost than o1, and it ran faster than any of the other models used (Roumeliotis et al., 2025).

3 Methods

3.1 Dataset and Preprocessing

We use the CFPB mortgage complaint dataset, restricted to records containing consumer narratives. Each complaint includes free-text describing the issue and a company response category. We define a binary risk_flag: complaints labeled “Closed with monetary relief” or “Closed with non-monetary relief” are treated as risky (1), and all others as non-risky (0). Only about 5% of complaints fall into the risky class, making accuracy misleading and motivating stratified splits and class-sensitive objectives. Narratives range from a few words to multi-page descriptions, consistent with patterns documented in prior studies

(Roumeliotis et al., 2025). We apply light text cleaning, lowercase conversion, PII removal, whitespace normalization, and preservation of domain-specific mortgage terminology (Bastani et al., 2019). Finally, we create train/validation/test splits using a 60/20/20 stratified partition to maintain class proportions across all subsets.
Code: https://github.com/gongbyung/ai-ethics-prototypes/blob/main/respect-cfpb/notebooks/01_eda.ipynb

3.2 Baseline model: TF-IDF with Logistic Regression

Our baseline model uses a TF–IDF bag-of-words representation paired with logistic regression. This reflects a standard, interpretable approach commonly used in earlier CFPB complaint analysis research (Bastani et al., 2019; Oyewola et al., 2023) and provides a strong comparison point for evaluating modern transformer models. We transformed each complaint narrative into a sparse TF–IDF vector using only unigram features. The resulting vectors were passed into a logistic regression classifier trained with class_weight="balanced" to compensate for the extreme class imbalance (~5% risky vs. ~95% non-risky), a pattern confirmed through our EDA. The model was trained on the stratified training split and evaluated on the held-out test set using accuracy, precision, recall, F1, and ROC-AUC, consistent with evaluation strategies used in prior work (Oyewola et al., 2023; Cui, 2025). Because accuracy is misleading under heavy imbalance, we emphasize class-1 (risky) precision/recall and overall ROC-AUC.
Code: https://github.com/gongbyung/ai-ethics-prototypes/blob/main/respect-cfpb/notebooks/02_modeling_nlp.ipynb

3.3 FinBERT Fine-Tuning

We selected FinBERT because domain-specific financial encoders consistently outperform general-purpose models on financial text classification (Vasudeva Raju et al., 2022; Oyewola et al., 2023), and CFPB complaints contain dense mortgage-, servicing-, and regulation-specific terminology.

Our first approach used LoRA adapters for parameter-efficient fine-tuning, combined with imbalance-mitigation techniques including class-weighted loss, balanced sampling, and multiple hyperparameter configurations. Despite its computational advantages (Hu et al., 2021), LoRA proved unstable under extreme class imbalance (≈5% risky): models frequently collapsed into predicting all non-risk cases or overcorrected toward the minority class as imbalance pressure increased.

These failures were systematic rather than random noise. Severe imbalance amplifies small optimization instabilities, pulling gradients toward the majority class (Oyewola et al., 2023; Cui, 2025). Risky and non-risky narratives are linguistically similar, often differing only in subtle procedural or regulatory cues (Roumeliotis et al., 2025). LoRA's low-rank parameterization struggled to learn these fine-grained distinctions, and small changes in rank, scaling, or dropout frequently flipped the model's behavior. Combining multiple imbalance strategies (loss weighting, oversampling, and balanced sampling) further destabilized training.

Because of this instability, we transitioned to full end-to-end fine-tuning of FinBERT. Unlike LoRA, which updates only low-rank adapters, full fine-tuning allows all parameters to adapt to the task distribution (Hu et al., 2021). This provided sufficient capacity for learning subtle, domain-specific distinctions between risky and non-risky narratives. Importantly, full FinBERT fine-tuning produced stable, reproducible results with only class-weighted cross-entropy loss, without requiring oversampling or balanced batching.
Code: https://github.com/gongbyung/ai-ethics-prototypes/blob/main/respect-cfpb/notebooks/02_modeling_nlp.ipynb

3.4 RoBERTa Fine-Tuning

We began another transformer experiment with the original BERT model as a baseline measurement of text classification to then build upon and fine-tune in an attempt to improve classification metrics. BERT was an adequate baseline model given that it is widely used as a benchmark for modern NLP tasks and performs relatively well on shorter text samples (Vasudeva et al., 2022). However, the CFPB complaint narratives contained high amounts of domain-specific data regarding regulatory and procedural language

that limited the traditional BERT model from being able to accurately predict the minority class in the dataset, given the very large initial imbalance between our risky and non-risky labels.

This initial model was trained using the standard BERT model on the original cleaned dataset and a maximum text length of 64 in order to limit the runtime and computational power needed to train the models. The results of this first model on the test set looked like an adequate baseline at first, but upon further investigation, it seemed like the model was performing very well on the non-risky class and failing to correctly predict the risky class in the vast majority of cases. This was a strong indication that the severe class imbalance was impacting the model's ability to effectively learn patterns in the data labeled as risky, and it would require some level of fine-tuning to address the issue.

In the next iteration of the model, we decided to train using RoBERTa in our transformer architecture instead of the baseline BERT model. Given the unique and domain-specific context of the dataset we were using in this experiment, we decided that RoBERTa would yield better results since it is trained on a far larger corpus of text and it optimizes the pretraining process through longer training times and dynamic masking. Research also indicates that RoBERTa is optimized for handling cases with severe class imbalance, making it a strong candidate for this study (Cui et al., 2025).

Our results for the first RoBERTa model demonstrated improvements in both the recall and f1-score for the minority risky class, while also generating small improvements in the ROC-AUC from the first BERT model. After further fine-tuning by decreasing the learning rate and increasing the maximum length of the input text used in training, we also decided to address the class imbalance by implementing a class-weighted cross-entropy loss as used in previous models to improve the training process for the minority class. Through these changes, our final version of the RoBERTa model yielded significant improvements over both the initial BERT and RoBERTa models. This is reflected by a higher precision in the minority class and the highest ROC-AUC of these three models.

Code: https://github.com/gongbyung/ai-ethics-prototypes/blob/main/respect-cfpb/notebooks/02_modeling_RoBERTa.ipynb

3.5 Post-Hoc Topic Modeling and Error Analysis

To diagnose the model's systematic errors, we conducted a post hoc topic modeling analysis of all true-risk complaints (TP + FN). We generated dense embeddings using all-MiniLM-L6-v2 and applied K-Means clustering (k = 14) to obtain semantically coherent groups. We interpreted each cluster by extracting representative n-grams from average TF–IDF weights and computed cluster-level recall to localize failure modes. This analysis revealed that low-recall clusters aligned with rule-based regulatory domains in which violations are defined by statute rather than narrative tone. This pipeline surfaced structural failure modes not visible from aggregate metrics alone.

Code: https://github.com/gongbyung/ai-ethics-prototypes/blob/main/respect-cfpb/notebooks/03_modeling_topic.ipynb

All preprocessing scripts, model training code, and experiment configurations are available at: https://github.com/gongbyung/ai-ethics-prototypes/tree/main/respect-cfpb

4 Results and Discussion

4.1 Baseline model: TF-IDF with Logistic Regression

The TF–IDF with logistic regression model provides a useful point of comparison for more advanced architectures. These results on the held-out test set reveal a characteristic pattern of imbalanced classification: the model achieves strong overall accuracy and a respectable ROC-AUC, indicating that TF–IDF features contain meaningful signal. However, the extremely low precision for the minority class shows that the model compensates for the imbalance by over-predicting the risky label. This behavior is clearly reflected in the confusion matrix, where 452 non-risky complaints are incorrectly labeled as risky, while 67 risky complaints are correctly identified. The baseline thus operates in a high-recall, low-precision regime for the risky class. While not a majority-class predictor, its reliance on surface-level lexical features limits its ability to differentiate between true regulatory harm and general expressions of

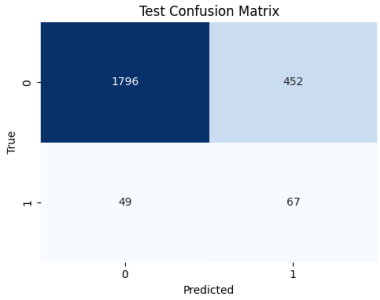
consumer dissatisfaction, motivating the need for more expressive models.

```
Test Results
              precision    recall  f1-score   support

           0      0.973     0.799     0.878      2248
           1      0.129     0.578     0.211       116

    accuracy                         0.788      2364
   macro avg      0.551     0.688     0.544      2364
weighted avg      0.932     0.788     0.845      2364

ROC-AUC: 0.7852228801079887
```



4.2 FinBERT Fine-Tuning

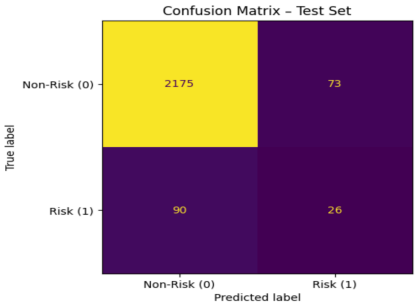
The fully fine-tuned FinBERT model demonstrates a clear performance improvement over traditional TF–IDF methods, particularly in its ability to assign calibrated risk scores. At this threshold, FinBERT attains 0.263 precision, 0.224 recall, and 0.242 F1 for the risky class, with overall accuracy of 0.931 and ROC–AUC of 0.777.

```
Using threshold = 0.448
              precision    recall  f1-score   support

           0      0.960     0.968     0.964      2248
           1      0.263     0.224     0.242       116

    accuracy                         0.931      2364
   macro avg      0.611     0.596     0.603      2364
weighted avg      0.926     0.931     0.928      2364

Test ROC-AUC: 0.777
```



These results show that full FinBERT fine-tuning produces substantially stronger overall performance than the baseline, especially in accuracy and minority-class precision. Although the ROC-AUC is similar to that of the TF–IDF model, the transformer exhibits more stable, better-calibrated decision boundaries. The confusion matrix illustrates this shift. The model generates far fewer false positives than the baseline (73 versus 452), while maintaining comparable, though lower, recall for the risky class. Consequently, FinBERT avoids the baseline's tendency to over-predict on risk and instead delivers more reliable, precision-oriented predictions. This reflects a meaningful improvement in distinguishing genuine regulatory harm from general consumer frustration, demonstrating the practical value of transformer-based fine-tuning for detecting nuanced risk signals in complaint narratives.

4.3 RoBERTa Fine-Tuning

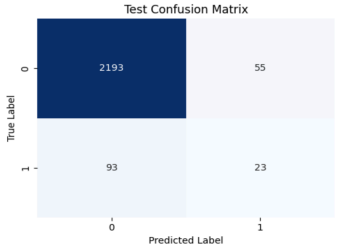
The final RoBERTa model demonstrates significant improvements in certain metrics over the baseline TF-IDF model and initial iterations of the BERT model that were then systematically fine-tuned. Our test metrics and confusion matrix for the final RoBERTa model are shown here:

```
              precision    recall  f1-score   support

           0      0.959     0.976     0.967      2248
           1      0.295     0.198     0.237       116

    accuracy                         0.937      2364
   macro avg      0.627     0.587     0.602      2364
weighted avg      0.927     0.937     0.932      2364

ROC-AUC: 0.7623193796784881
```



The most notable change is the increase in precision for the risky class, from 0.129 in the baseline to 0.295. This likely reflects RoBERTa's ability to capture deeper semantic patterns than the TF–IDF model, which relies only on word counts. However, as the precision increases in the RoBERTa model, this also forces the recall to drop as the model becomes more conservative in its prediction of the risky class, leading to less frequent positive predictions. The overly conservative approach of the RoBERTa classification also contributes to the model's slightly decreased ROC-AUC score as performance becomes weaker across different classification thresholds.

Due to the tradeoff between precision and recall, another useful metric in evaluating the model improvement is the f1-score, which can help determine the accuracy and consistency of the models in identifying the minority class, as it considers both precision and recall. The RoBERTa model demonstrates an increase in the f1-score over the TF-IDF model from 0.211 to 0.237, indicating that overall RoBERTa may be more suitable for real-world applications of new unseen data that is highly imbalanced. This result is consistent with prior research that has shown RoBERTa's lower sensitivity to class imbalance and improved precision for the minority class in a similar context (Cui et al., 2025).

4.4 Post-Hoc Topic Modeling and Error Analysis

Model analysis reveals a consistent pattern across the model's decisions. High-confidence false positives contain strong harm-oriented or regulatory language, making them linguistically similar to true risk cases. False negatives, however, tend to describe clear procedural or statutory violations, such as escrow misapplication, flood-insurance handling, or credit-reporting errors, but in calm, administrative language. Because these violations hinge on regulatory rules rather than narrative tone, the classifier frequently under-detects them. Topic modeling corroborates this: the lowest-recall clusters concentrate in rule-based regulatory domains, where risk is structural rather than linguistic. Overall, the model is sensitive to harm-signaling language but less effective at identifying violations expressed without strong sentiment cues.

5 Conclusion

This study shows that fine-tuned transformer models measurably improve classification of risky and non-risky CFPB complaints. By implementing models that are trained on larger and more domain-specialized corpora, our improved models were better suited for capturing complex semantic meanings in consumer narratives than the baseline TF-IDF model that largely classified examples based on word frequency. The research conducted also highlights the limitations caused by significant class imbalance and effective approaches that can be applied to mitigate this. These context-specific imbalances are important to consider for users who may decide to implement similar algorithms at the CFPB.

Despite the improvements made over the baseline model in both precision and f1-score, there is still significant work that needs to be done in order to increase the classification accuracy of the minority risky class to a point where these models can be used reliably to completely automate risk classification. Future work may explore alternative loss functions for severe imbalance, additional transformer variants, and retrieval-augmented architectures.

Another practical next step is to incorporate a Retrieval-Augmented Generation (RAG) layer. By retrieving relevant Consumer Financial Protection Laws and Regulations, investor guidelines, and internal policies, the system could check whether the facts described in a complaint conflict with a specific rule, even when the narrative tone is neutral. This provides a complementary capability that the classifier alone cannot capture, helping to improve recall on rule-driven violations without sacrificing precision.

Appendix

Contributions

Alex Kim: This project was a collaborative effort. I contributed to the exploratory data analysis, dataset preprocessing, and development of class-imbalance mitigation strategies. I implemented the baseline TF–IDF + logistic regression model, the LoRA fine-tuning experiments, and the full FinBERT fine-tuning pipeline. I also designed and conducted post-hoc topic modeling and error analysis to diagnose systematic model failures. In addition to technical work, I contributed to the Introduction, literature review, and the Methods and Results sections.

Patrick Abousleiman: This project was closely coordinated among our team. I worked on some of the initial exploratory data analysis, looking into and visualizing the distributions of the dataset. I implemented the BERT, ModernBERT, and RoBERTa models before fine-tuning further iterations of the RoBERTa model to improve accuracy metrics. Beyond the modeling process, I contributed to the background, methods, results and discussion, and conclusion sections of the final report.

References

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems With Applications, 127, 256–271. https://doi.org/10.1016/j.eswa.2019.03.001

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. In Proceedings of the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML '17).

Cui, X. (2025). Addressing data imbalance in transformer-based multi-label emotion detection with weighted loss. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025).

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LORA: Low-rank adaptation of large language models (Version 2). arXiv.

Krause, D. (2025). Dismantling financial oversight: Implications of CFPB downsizing for regulatory integrity and market stability. SSRN.

Oyewola, D. O., Omotehinwa, T. O., & Dada, E. G. (2023). Consumer complaints of consumer financial protection bureau via two-stage residual one-dimensional convolutional neural network (TSR1DCNN). Data and Information Management, 7, 100046. https://doi.org/10.1016/j.dim.2023.100046

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Think before you classify: The rise of reasoning large language models for consumer complaint detection and classification. Electronics, 14(6), 1070. https://doi.org/10.3390/electronics14061070

Vasudeva Raju, S., Bolla, B. K., Nayak, D. K., & Jyothsna, K. H. (2022). Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings. In Proceedings of the 2022 IEEE 7th International Conference for Convergence in Technology. I2CT.