Exploring Banking Complaint Patterns and Outliers with Deep Generative Models

Jonathan Hague
jhague@stanford.edu

Abstract

Financial institutions in the United States are regulated and supervised by multiple federal agencies, including the Consumer Financial Protection Bureau (CFPB), which is tasked with "protecting consumers against unfair, deceptive, or abusive practices" [1]. Since its inception, the CFPB levied billions in penalties and secured billions in relief to 195 million consumers. Consumer complaints lodged with the CFPB are critical for identifying issues that prompt investigations and enforcement actions. Identifying themes from these complaints can be useful for financial institutions to mitigate regulatory risk and proactively address emerging consumer issues.

This paper explores topic classification as a downstream task of XLNet (Yang et al. 2019), a generative pretrained model that has demonstrated text classification performance (Minaee et al. 2020), by working with text embeddings of a select CFPB complaints dataset to derive themes and outliers. We evaluate various clustering methods on our lower dimension embeddings to identify outliers. Subsequently, we treat each cluster of relevant complaints as its own concatenated document and apply traditional TF-IDF on those clusters (now documents) to derive themes.

Our architecture[2] identified two insightful themes in the data: complaints about Consumer Facing Services, and complaints about International and Complex Transactions. Additionally, outlier detection revealed what appears to be a spam attack on the CFPB website targeting major US banks in a systemic manner. These findings highlight the potential of generative pretrained models, combined with advanced machine learning techniques, in enhancing regulatory compliance and improving consumer care efforts in the financial industry.

1 Introduction

The primary issue with analyzing complaint texts lies in their high dimensionality and complex language. Each complaint carries unique semantic signatures that are difficult to quantify with basic tools. Current methods like keyword searches and basic statistical techniques often fail to capture the essence of complaints, missing contextual nuances and semantic relationships.

Effective analysis of complaint data is critical for businesses, leading to improvements in service strategies, product development, and customer retention. Identifying trending themes proactively can help financial institutions avoid severe enforcement actions and save substantial costs. Models like XLNet succeed by capturing deeper linguistic contexts and maintaining robustness over varied inputs but require substantial computational resources and can be opaque.

This research uses a text analysis pipeline with XLNet for embeddings, followed by dimensionality reduction and clustering. The text is transformed into a high-dimensional space with XLNet to capture nuanced meanings, then made manageable with dimensionality reduction while preserving essential characteristics. Clustering algorithms group embeddings into meaningful clusters, which are further analyzed with TF-IDF (term-frequency inverse document-frequency) to identify distinct themes. This architecture allows real-time identification of emerging complaint themes.

[1] https://www.consumerfinance.gov/about-us/
[2] https://github.com/Jonathan-Hague/CS224N

2 Related Work

Analyzing consumer complaints has evolved from traditional keyword extraction and basic statistical techniques, which lacked deep semantic understanding, to advanced natural language processing (NLP) models like BERT and GPT that excel in text classification and generation. While BERT's bidirectional training improves context understanding, and GPT's autoregressive model generates coherent text, both struggle with longer sequences. Transformer-XL addresses this by extending the context window, enhancing long-range dependency capture, a strength further refined by XLNet through permutation-based training for bidirectional context utilization. Dimensionality reduction techniques like t-SNE and UMAP, along with clustering methods such as K-means, DBSCAN, and HDBSCAN, have facilitated effective visualization and grouping of high-dimensional data. In the financial sector, research leveraging these techniques has primarily focused on sentiment analysis and fraud detection, with limited studies on comprehensive analysis of CFPB complaints. Our work extends this by using XLNet for embeddings, UMAP for dimensionality reduction, and K-means for clustering to derive actionable themes from consumer complaints, addressing a critical literature gap. This integrated approach enhances thematic analysis accuracy and scalability, contributing to improved regulatory compliance and consumer protection in the financial industry.

3 Approach

Transformers typically have a fixed context, which restricts their ability to retain earlier data while predicting future data. To address this limitation, Transformer XL proposes a method to extend the context window of Transformers, ensuring the model does not forget previous data. This extension is crucial for tasks like language modeling, where predicting the next word x_t given all previous words $x_{<t}$ is essential. Where given a corpus of tokens $x = (x_1, x_2, \ldots, x_T)$, the aim is to model the probability distribution $p(x) = \prod_t p(x_t \mid x_{<t})$. The context $x_{<t}$ is encoded into a fixed-size hidden state, which the Transformer uses to predict the next token.

3.1 Extending Context with Transformer-XL

Transformer-XL (Dai et al. 2019) introduces an architecture that extends the context of Transformer models during both training and evaluation. The extent of the context is controlled by a hyperparameter M, which determines the length of the context window. While a larger M provides a greater context, it also increases computational costs.

The model enhances the context by incorporating concepts from Recurrent Neural Networks (RNNs). Specifically, we concatenate the hidden layers h of the current segment with the hidden layers from previous segments. This approach allows us to create queries (Q) from the new segment, while keys (K) and values (V) can originate from either the new or old segments.

During the forward pass in training, this method increases the context length. However, backpropagation can encounter issues such as memory limitations and vanishing gradients. To mitigate these problems, we truncate the gradient computation to a limited context, effectively stopping backpropagation beyond a certain point.

3.2 Relative Positional Encoding

The extended context requires a shift from absolute to relative positional encoding. Absolute positional encoding is impractical for longer sequences because it assigns a unique vector to each position, leading to large and unmanageable matrices. This encoding also fails to distinguish between positions effectively when the sequence length exceeds a certain limit.

Instead, Transformers XL uses relative positional encoding, which encodes the relative distance between word pairs. This approach ensures that the positional encoding matrix does not grow excessively large, maintaining computational efficiency.

By adopting these strategies, Transformer-XL can process longer sequences during evaluation, with the length of the sequence determined by the number of hidden states in the transformer. This approach significantly improves the model's ability to capture long-range dependencies, enhancing its performance in language modeling tasks.

3.3 XLNet

To further increase the context length, XLNet combines the Autoregressive and Autoenconding language modeling. By combining these techniques, XLNet extends the context length, improves efficiency, and enhances language modeling capabilities, offering significant advantages over traditional models.

3.3.1 Autoregressive Language Modeling (AR LM)

AR LM aims to predict the next word given the context. For a corpus X, this involves predicting the next word in the forward pass and the previous word in the backward pass, expanding probabilities using the chain rule. This bidirectional context sets it apart from unidirectional models like GPT and masked language models like BERT.

Given a corpus $X = (x_1, x_2, \ldots, x_T)$, AR LM models the probability $p(x) = \prod_t p(x_t \mid x_{<t})$. In Transformer-XL, this is achieved through permutation-based pre-training, allowing the model to learn bidirectional context by considering all permutations of input tokens. This permutation-based approach is also reflected in the SentencePiece tokenizer (Kudo and Richardson 2018) used by XLNet.

In the forward pass, we parameterize the first term and expand it to model the probability. The log of the product becomes the summation of logs, and we model the forward pass with a transformer decoder, multiplying the word embeddings at the target and dividing by the softmax of the word embeddings at the targets.

The attention masks are randomly sampled based on permutations, allowing the model to account for all possible permutations. However, this is computationally expensive, as it requires summing over all probabilities. To address this, we limit the permutations and predict only a portion of the sentence. We choose a constant C based on the sentence length (e.g., 10% of the sentence). Thus, the equation no longer includes a summation, and we only predict words occurring after index C. This reduces computational costs, making the loss function more efficient.

Permutations are feasible because, during prediction, we maintain the order while altering the masking. We always predict in order, ensuring that location Z_t is not forgotten. While predicting word x_{Z_t}, we can focus on the location but not the word vector itself. To achieve this, we use two attention heads: one for the content stream and one for the query stream.

This dual-stream mechanism is essential as transformers process entire sentences in parallel rather than word-by-word. The query stream mask excludes the diagonal, ensuring it does not attend to the content. Predictions utilize the output from the query stream.

3.3.2 Autoencoding Language Modeling (AE LM)

AE LM involves corrupting the sentence and masking 15% of the words, then predicting these masked tokens. We compute the probability P as the maximum log probability of the masked token conditioned on the corrupted sentence. This loss function applies only to masked tokens, as $m_t = 1$ indicates a masked token.

For AE LM, we assume the independence of outputted words given the corruption, which simplifies the modeling but introduces a strong assumption. This probability is modeled using the transformer's encoder rather than the decoder, focusing on the T-th word in the sentence as output by the encoder.

During pre-training, masked tokens are present, creating a discrepancy during fine-tuning when no masked tokens exist. This bias can affect the model's performance. XLNet addresses this by generalizing the AE LM (BERT) assumption. Instead of assuming independence, XLNet relaxes this assumption enhancing context and overall model performance.

By combining these techniques, XLNet extends the context length, improves efficiency, and enhances language modeling capabilities, offering significant advantages over traditional models.

3.3.3 Tokenization with SentencePiece in XLNet

XLNet uses SentencePiece model for tokenization, differentiating itself from BERT and GPT models, which use WordPiece and Byte Pair Encoding (BPE) respectively. SentencePiece treats input as raw

bytes, making it language-independent and capable of handling out-of-vocabulary tokens by breaking words into subwords, balancing word-level and character-level tokenization.

Given the diverse educational backgrounds and writing styles of consumers entering complaint texts on the CFPB website, we needed a tokenizer minimally sensitive to orthographic variations. SentencePiece was ideal for this purpose. Using the cased variant, we retained the case sensitivity of input text to capture the semantic emphasis of uppercase letters, hence employing the 'xlnet-large-cased' variant.

SentencePiece consists of four components: the normalizer, the trainer, the encoder, and the decoder. The normalizer converts semantically equivalent Unicode characters into canonical forms using Unicode NFKC normalization. The trainer analyzes the text corpus to train the tokenizer and detokenizer. The encoder translates sentences into sequences of integers, while the decoder performs the reverse operation. Each token is mapped to a unique integer ID for processing by the model.

SentencePiece performs subword segmentation using methods like byte-pair encoding (BPE) or the Unigram Language Model, training directly from raw sentences without pre-tokenization. This ensures lossless tokenization, treating input text as a sequence of Unicode characters, adhering to the formula: `Decoder(Encoder(Normalize(Text))) = Normalize(Text)` . It replaces white spaces with a meta symbol _ (U+2581), making it efficient for languages without white spaces.

Under the Unigram Language Model, a sentence can have multiple segmentation candidates, treated as subwords. The probability of the entire subword sequence is the product of each subword's probabilities, which sum to 1. Hidden variables $P(x_i)$ given the vocabulary are estimated by maximizing the marginal likelihood using the Estimation Maximization algorithm.

3.3.4 Dimensionality Reduction, Clustering, and Theme Extraction

To analyze the embeddings, we utilized Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. UMAP maps high-dimensional data into a lower-dimensional space (2D) while preserving the significant structure. This allows close points in high-dimensional space to remain close in 2D. Each sentence, originally a high-dimensional vector (e.g., 1024 dimensions with XLNet-Large), is transformed into two coordinates that encapsulate the essence of the sentence's meaning or structure. The 2D UMAP embeddings were visualized in Figure 1 below.

We explored various clustering techniques before settling on K-means. We tried t-SNE, DBSCAN, and HDBSCAN clustering. K-means had the best representation of our lower dimension embeddings. We used the Elbow Method to determine the optimal number of clusters by plotting inertia values for a range of K values as shown below in Figure 2. This allowed us to identify $k = 2$ as the optimal K value.

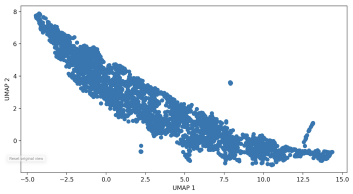

Figure 1: UMAP Projection of the Embeddings

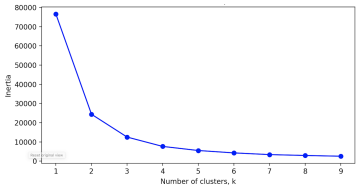

Figure 2: Elbow Method for Optimal K

We then plotted the clustered distribution of the 2D embeddings in Figure 3 for visual inspection.

We noticed in Figure 3 data points that were far from the clustered points which seemed like outliers. We then applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect those outliers. We set the maximum distance (ϵ) to 0.5 and the minimum number of samples to 10. This method helped identify and visualize outliers in the dataset as shown in cluster 1 and 2 in Figure 4 below.

Finally, we used TF-IDF on the associated complaints within each of the two clusters to indetify themes. To acheieve this task, we first grouped the complaints by clusters and concatenated the texts within each cluster (the two clusters are now two documents). We then converted these texts using CountVectorizer into a matrix of token counts. We filtered out stopwords and terms with document

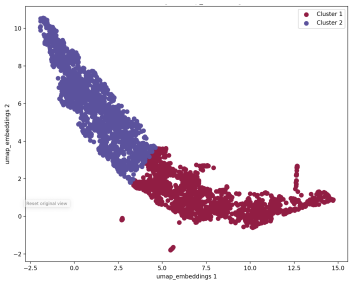

Figure 3: K-Means Clustering of 2D Embeddings With K=2

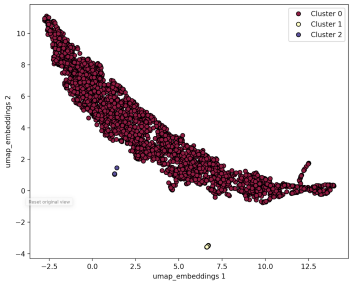

Figure 4: DBSCAN Clustering of 2D Embeddings for Outlier Detection

frequency above 40%. TfidfTransformer was then applied to compute TF-IDF scores wich quantifies the importance of each term within and between the clusters.

Our baseline for theme comparison was a regular TF-IDF method applied to the corpurs of complaints that allowed us to extract themes and their related complaint document using Non-negative Matrix Factorization (NMF).

4 Experiments

4.1 Data

The data consists of consumer complaints filed with the CFPB[3] between January 1, 2019 and December 31, 2019 against Bank of America National Association. These are 3,007 complaints filed from 52 state, district, and territory covering banking services. For the purpose of our project, we retained only the text of the complaint and each complaint respective ID number.

4.2 Evaluation method

We used regular TF-IDF method on our complaint data as a baseline to compare against derived themes in our proposed architecture. This baseline method produced two themes. We inspected for each theme the 10 words with highest TF-IDF score to qualitatively derive themes based on our subject matter expertise in the field. This baseline method produced a theme around Disputed Transactions, and another theme around Checking Account Transactions. The first theme was derived from the following top ten words with highest TF-IDF scores: payment, claim, fraud, report, information, charge, late, balance, loan, letter. The second theme was derived from the following top ten words with the highest TF-IDF scores: check, funds, hold, deposit, available, business, days, cash, branch, said.

Our proposed architecture produced two themes derived from the following two clusters: cluster 1 with the following top ten words with highest TF-IDF scores: withdrawal, affiant, purchaser, reconciled, margin, pretences, yelling, transit, medallion, stamps. Cluster 1 theme was assigned as complaints around Consumer Facing Services. Cluster 2 provided the following top ten words with the highest TF-IDF scores: international, master, citizenship, irma, inherited, wealth, deficiency, redemption, instruments, reviews. Cluster 2 theme was assigned as complaints around International and Complex Transactions.

To evaluate both approaches, we selected a random sample of 50 complaints (representing around 1.7% of our data) and compared the accuracy, precision, and recall on themes assigned by each method. True values (themes) were assigned by us for the baseline and for the proposed architecture after evaluating each complaint and its relevance to one of the two themese identified by each method. True Positive means the complaints belongs to Theme 1, and True Negative means the complaint belongs to Theme 2. We then calculated accuracy, recall, and precision on this sample, and compared the performance of our architecture against that of the baseline TF-IDF.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

[3]https://www.consumerfinance.gov/data-research/consumer-complaints/

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$

4.3 Experimental details

Our baseline method was run on CPU and was computationally effective. It produced all the results in less than 10 minutes.

Our proposed architecture required an NVIDIA Tesla GPU on Google Cloud to run the embeddings from XLNet on our dataset. This took less than an hour to compute. Once the embeddings were computed, we then downloaded them and used those locally on CPU. The rest of the exercise took less than 10 minutes to complete. The locally used CPU is an Apple MacBook Pro 2020, Core i5, 16 GB RAM.

4.4 Results

Detected outliers have been identified in this project as spam attacks targeting major financial institutions that do not seem to have been filtered out by the CFPB. We traced one of the outliers to one complaint that propagated with slight textual changes throughout multiple financial institution. These are taxing spam attacks as financial institutions are required by law to respond or resolve each complaint within 15 calendar days. Unfiltered spam attacks on CFPB put unnecessary strain on the financial sector.

We traced one significant outlier to complaint number 3421336 shown in Figure 5. This was submitted 505 times with slight textual variations targeting 108 financial institutions between 2019 and 2020 as shown in Figure 6. This outlier was identified in this paper as a targeted spam that was not filtered out by the CFPB.

Consumer complaint narrative

This particular account situation that is lately filing on my own credit document has a seriously unfavorable relation to my personal ability to obtain a present loan application. I highly recommend you generate verification that XXXX XXXX has been reported completely in accordance with the Fair Credit Reporting Act regulations, it's really a serious problem to misreport. More confirmation of the aforesaid item too. My proper request must over, I was never 30 days/60 days/120 days late in any of my payments and I'm not greatly tuned in to the date opened so I prefer to ask you be investigated as soon as possible and confirmed to be correct. Thanks!

Figure 5: CFPB Consumer Complaint Number 3421336

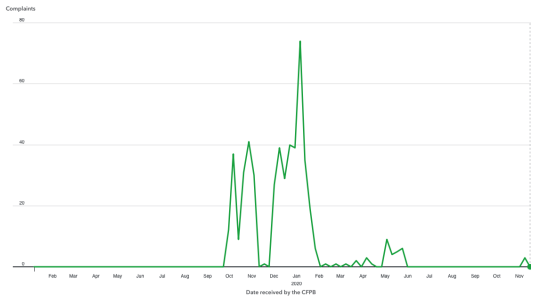

Figure 6: Frequency of Spam Attack Traced to Complaint Number 3421336 Between 2019-2020

The baseline method produced an accuracy on the randomly selected sample of around 38%, and a precision and recall of around 40% each (Figure 7). TF-IDF performed poorly on its task of deriving thematic analysis of our sample.

Our proposed architecture, on the other hand, performed really well on the randomly selected sample. Accuracy reached 62%, recall and precision reached around 60% each (Figure 8). Our proposed architecture almost doubled the accuracy of identifying and assigning individual complaints to accurate themes.

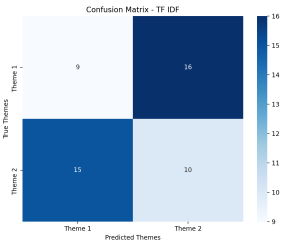

Figure 7: Confusion Matrix of Baseline TF-IDF

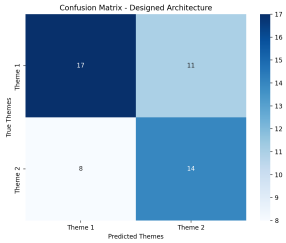

Figure 8: Confusion Matrix of Designed Architecture

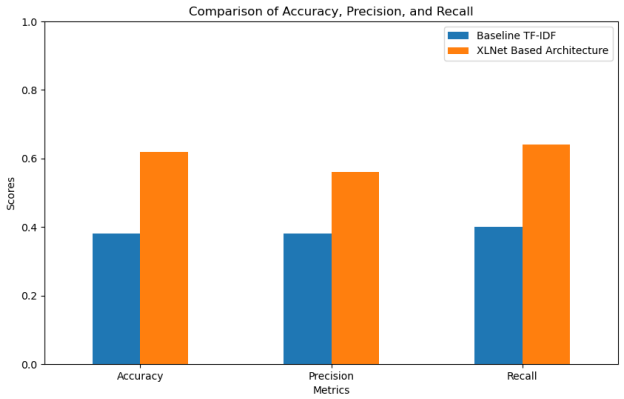

Figure 9: Metric Comparison Baseline Vs Proposed Architecture

Overall, using XLNet allowed our thematic analysis of consumer complaints to capture more context and better theme identification. This, in turn, allowed for a better accuracy, recall, and precision (Figure 9).

These results make sense given XLNet high performance on text classification tasks, which has translated here in powerful embeddings that allowed for better capture of themes in our data.

5 Analysis

Generative pretrained models are powerful tools to generate output. But the usage of these models latent spaces are equally important when it comes to downstream tasks in specialized fields where a deep understanding of the domain is as important as the ability to understand and work with deep learning models.

In our proposed architecture, we used the lower dimensional embeddings of our data to divide our corpus into clusters or documents. That step allowed us to have two documents that contain more meaning than when directly applying a bag of words method directly on the text itself. Once we identified those documents (or clusters), we then reverted to TF-IDF and forced the method to treat the identified clusters as documents.

This allowed our accuracy to double on the random sample we used to evaluate our method against baseline. For future work, it would be useful to boost the evaluation on the overall data by allowing another model to evaluate the classification of the complaints into the two derived themes (for baseline and proposed architecture). However, human inspection and evaluation would still be an essential part of this work given the required domain knowledge expertise.

6 Conclusion

This paper demonstrated the efficacy of XLNet for thematic analysis of consumer complaints filed with the CFPB. By leveraging XLNet's contextual embeddings with dimensionality reduction and clustering techniques, we significantly improved theme identification accuracy compared to traditional TF-IDF methods. Our approach identified two primary themes (complaints about Consumer Facing Services, and complaints about International and Complex Transactions) and revealed a systematic spam attack targeting major US banks. These findings highlight the potential of integrating generative pretrained models with domain-specific expertise for real-time monitoring and analysis of consumer complaints, providing financial institutions with actionable insights to mitigate regulatory risks and enhance consumer care. Future work would include expanding the dataset, exploring additional machine learning methods, and incorporate human-in-the-loop feedback to further refine theme detection. This would be to advance thematic complaint analysis systems and improve regulatory compliance and consumer protection in the financial industry.

7 Ethics Statement

The potential for model bias and unfair treatment of certain consumer groups is addressed by auditing datasets, using fairness-aware algorithms, and involving diverse review teams. Transparency and explainability challenges, due to the complex nature of models like XLNet, are tackled by developing interpretable models and documenting decision-making processes. These strategies collectively ensure the responsible development and deployment of CFPB complaints analysis technology.

References

Dai, Zihang et al. (Jan. 2019). "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context". In: *arXiv e-prints*, arXiv:1901.02860, arXiv:1901.02860. DOI: 10.48550/arXiv.1901.02860. arXiv: 1901.02860 [cs.LG].

Kudo, Taku and John Richardson (Aug. 2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *arXiv e-prints*, arXiv:1808.06226, arXiv:1808.06226. DOI: 10.48550/arXiv.1808.06226. arXiv: 1808.06226 [cs.CL].

Minaee, Shervin et al. (Apr. 2020). "Deep Learning Based Text Classification: A Comprehensive Review". In: *arXiv e-prints*, arXiv:2004.03705, arXiv:2004.03705. DOI: 10.48550/arXiv.2004.03705. arXiv: 2004.03705 [cs.CL].

Yang, Zhilin et al. (June 2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *arXiv e-prints*, arXiv:1906.08237, arXiv:1906.08237. DOI: 10.48550/arXiv.1906.08237. arXiv: 1906.08237 [cs.CL].