Team 15 Project

Consumer Complaint Segmentation: Automated Product & Issue Identification

GitHub Link: https://github.com/mahesh973/TagMyComplaint

Team Members: *Mahesh Babu Kommalapati, Shivani Ashish Mundle, Sharanya Senthil*

Introduction

Consumers often face issues or have complaints regarding financial products and services offered by banks, lenders, and other institutions. When these issues cannot be resolved directly with the company, consumers can turn to the Consumer Financial Protection Bureau (CFPB). CFPB is a U.S. government agency which ensures fair and transparent practices in the financial sector. However, consumers frequently struggle to accurately categorize their complaints into the correct product, issue, and sub-issue categories when filing a complaint through the CFPB's portal. This is due to a lack of understanding of the various financial products and a failure to identify the specific issue they are encountering. This manual categorization of complaints by the consumers is error prone. This can lead to misrouted complaints and inefficient resolution processes, as complaints may end up with the wrong department or team, resulting in delays and frustration for both consumers and companies.

Our motivation is to provide consumers with an automated tagging system that accurately identifies the product, sub-product, issue, and sub-issue categories based solely on their narrative description of the complaint. [Fig 1]

We also want to provide some insights on the consumer problems, the issue resolution process and how it can be made more efficient. By enabling consumers to merely explain their issue through a narrative, without having to navigate complex product and issue categorizations, we aim to save time and effort while enhancing customer satisfaction. This application can be extended to any consumer complaint domain, such as e-commerce, finance, or other service industries, where customers can simply describe their problem, and the automated system will handle the precise categorization and routing of the complaint.

Figure 1: Complaint Narrative classification application.

Methodology

Low Risk:

The low risks in our project included the potential oversimplification of categories by combining related products and the presence of inconsistencies in consumer-labeled categorization. To mitigate these risks, we performed thorough Exploratory Data Analysis (EDA) [Fig 2] [Fig 3] to understand the data balance across products, sub-products, issues, and sub-issues. This analysis guided our category grouping strategy, ensuring that we retained important information while adding more meaningful data points. The grouping process also combined similar complaints that were previously categorized inconsistently, resulting in a more coherent dataset. We iteratively refined our categorization strategy based on validation results, in alignment with the financial domain knowledge.

By thoroughly understanding the data and strategically grouping categories based on EDA insights, we mitigated the risks associated with oversimplification and inconsistent labeling, ultimately enhancing the quality and reliability of our models.

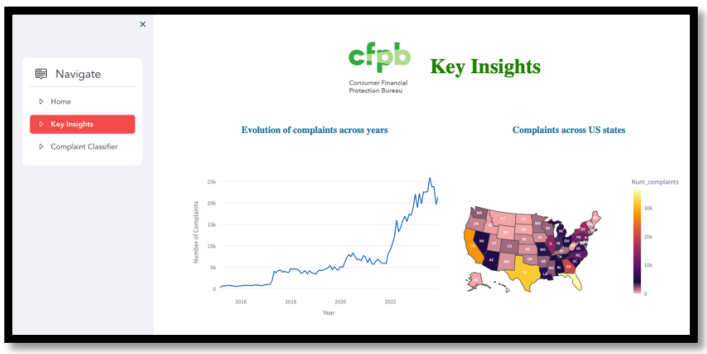

Figure 2: Key Insights for top 5 issues in a product states.

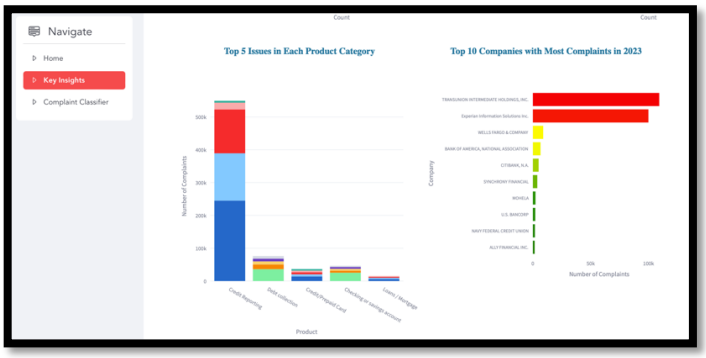

Figure 3: Key Insights for complaints trend across years and & top 10 companies with most complaints.

Medium Risk:

The medium risk in the project involved finding a balance between traditional machine learning methods and advanced transformer models for complaint categorization. To mitigate this, a hybrid approach [Fig 4] was adopted: transformers were utilized for predicting product and issue categories due to their superior performance, while traditional machine learning models were employed for refining predictions on smaller subcategories.

This strategy made use of traditional models because they're easier to understand and work with, especially when dealing with each product and issue separately. By developing separate models for each product and issue category, tailored to their specific characteristics, the project ensured robust classification of sub-products within each product category and sub-issues within each issue category. This approach capitalized on the strengths of traditional machine learning methods, allowing for accurate classification while mitigating the computational demands associated with transformer models. Additionally, a pipeline was established to predict issues and products first, followed by the application of

specific models for finer classification into sub-issues and sub-products. This streamlined approach optimized the efficiency of the categorization process while maintaining high accuracy levels across all levels of categorization.

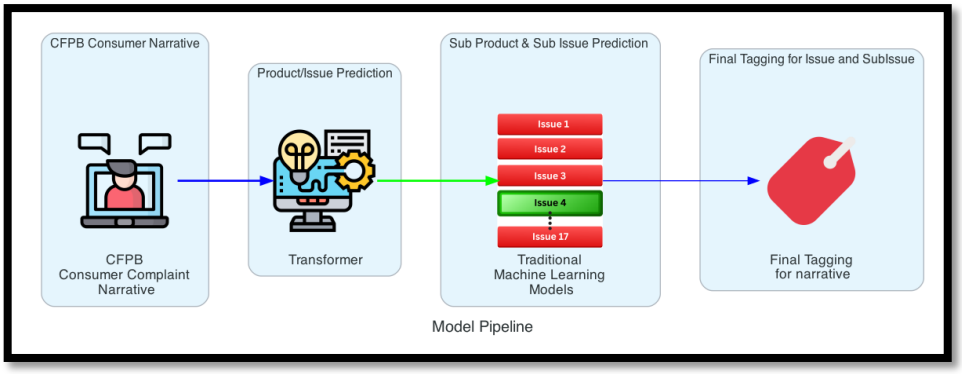

Figure 4: Pipeline to show the use of transformer and traditional machine learning models

High Risk:

The high risk we highlighted in proposal was developing a single pipeline for predicting various categories, prone to compounding errors and data imbalance issues. To mitigate the high risk, we developed two separate pipelines for product and issue categorization. [Fig 5] [Fig 6] This approach ensured that the prediction of issues remained independent of product categorization, thus reducing the risk of compounding errors. [Table 1] We conducted thorough testing to validate the effectiveness of this approach. Additionally, when faced with the challenge of transformer models failing with sub-product and sub-issue classification due to limited data, we implemented traditional classification models for each product and issue. This allowed for more accurate classification of sub-categories within the larger categories. To address data imbalance issues, we included data from previous years, ensuring a more balanced dataset where possible. Developing these models in parallel enhanced efficiency and scalability. While there remains a potential scenario where the issue and product classifications may diverge, the risk of error propagation has been significantly mitigated through this approach.

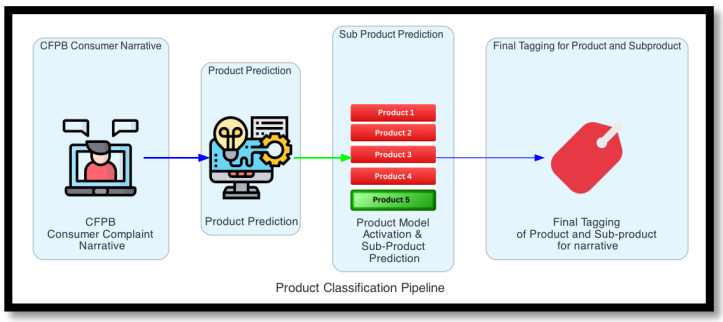

Figure 5: Issue Classification Pipeline

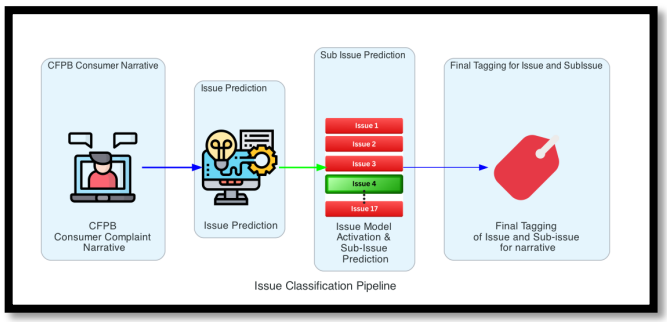

Figure 6: Product Classification Pipeline

Sr No	Category	Model Used	Weighted Average Precision	Accuracy
1	Product	Transformer (DistilBERT)	0.88	0.96
2	Sub Product	Random Forest Classifier	0.60	0.93
3	Issue	Transformer (DistilBERT)	0.72	0.77
4	Sub Issue	Random Forest Classifier	0.62	0.75

Table 1: Model Performance for each category

Conclusion:

In conclusion, our project "Consumer Complaint Segmentation: Automated Product & Issue Identification" has successfully developed an advanced system for automating the categorization of consumer complaints in the financial sector. By employing a hybrid approach, and that integrates transformer models with traditional machine learning techniques, we have achieved accurate classification of complaints into their respective product, sub-product, issue, and sub-issue categories. The implementation of separate pipelines for product and issue categorization effectively mitigated the risk of compounding errors and ensured the independence of predictions. Our solution enhances the efficiency of the complaint resolution process and improves customer satisfaction by providing a streamlined platform for consumers to express their concerns without requiring intricate categorization knowledge. This project exemplifies the potential of machine learning and language models in enhancing the accuracy and efficiency of complaint handling in the financial industry.

References:

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