CONSUMER COMPLAINTS CLASSIFICATION

Surya Theja Dokka, Pravalika Papasani, Manogna Lakkadasu, Vemulapalli Sri Samadarsini, Vineeth Nunna

Abstract

In this project, we developed a Natural Language Processing (NLP) system to automate the classification of consumer complaints in the financial sector. We applied several machine learning models—Multinomial Naive Bayes, Random Forest, Decision Tree, Gradient Boost, XGBoost, and KNN—along with the advanced NLP model BERT (Bidirectional Encoder Representations from Transformers), aiming to improve the efficiency of complaint resolution processes. Using data from the Consumer Financial Protection Bureau, we performed extensive preprocessing to prepare for effective model training. Each model was evaluated on metrics such as accuracy, precision, recall, and F1-score. XGBoost showed the best overall performance, while BERT excelled in handling complex textual content, and KNN demonstrated high precision but faced challenges in model generalization. This project highlights the utility of NLP and machine learning in streamlining complaint management in the financial sector, which could lead to quicker resolutions and enhanced customer satisfaction.

1 Project Description

Our project is creating a Natural Language Processing model that helps consumers easily report problems they're having with their bank, credit card, or loan. Currently, if you have a problem and want to complain, you must choose the correct category for your complaint from a long list. This can be confusing because the categories are full of financial terms that are not easily understandable by everyone.

Our project is designed to solve this problem. It works by reading the description of your issue that you type in, and then it automatically suggests the right category for you. This means you don't have to worry about understanding all the complex financial terms (categories) just to make a complaint.

We decided to work on this project because we noticed that choosing the wrong category can cause delay in complaint resolution. If your complaint ends up in the wrong place, it can take a lot longer to get sorted out. With our tool, we hope to make the whole process faster and easier for everyone.

By making it simpler to file complaints in the right category, our project aims to speed up the time it takes for problems to get resolved. This not only makes customers happier because their issues are fixed faster, but also helps banks and financial companies to manage complaints more efficiently. Our ultimate goal is to improve the way of handling complaints, making the financial world a bit easier for everyone to navigate.

2 Related Work

With the progress in our project of developing a Natural Language Processing (NLP) model for consumer complaint classification in the financial sector, we explored various previous research that aligns with our project objectives and methodologies.

Thomas [1], demonstrated efficient LSTM-based automated tools for classifying consumer complaints on digital platforms. They used deep learning models to achieve high accuracy which helps us in our approach to data pre-processing. We used similar techniques like stop words removal, tokenization and other data cleaning steps so that our data will be prepared for model training to lay a solid foundation to our consumer complaint classification system.

The research done by the authors in [2] included insights of performance of the pre-trained language models like BERT [5] while comparing it to the traditional machine learning models like TF-IDF. From the results of this comparison, we chose to utilize the BERT [5] model to fine-tune our training dataset for the classification.

Pramod Kumar Naik et al. [3] explored various machine learning and deep learning algorithms in consumer complaint classification and stated the advantage of using best algorithms, data preprocessing and model evaluation. From their insights we designed our strategy to evaluate different models like Naive Bayes, Random Forest, Decision Tree, Gradient Boost, and XGBoost to make sure that our approach will align with the objective of classifying consumer complaints in the financial domain.

Furthermore, the study of Bozyigit, F., Dogan[4] explained how machine learning approaches can be used in classification of complaints in the food industry. Although their focus was on the food industry, we can adapt some of their methodologies in the financial sector. We chose the XGBoost model to be included in our classification as it provided high accuracy in their classification of consumer complaints in the food industry. So, we would like to explore how it works with our data set in the finance sector.

3 Methods

The workflow for our consumer complaints classification project involves several key steps aimed at preparing the data, building baseline models(Multinomial Naive Bayes, Random Forest, Decision Tree, Gradient Boosting, and XGBoost), and evaluating the performance of a pre-trained language model named Bidirectional Encoder Representation from Transformers (BERT) [5]. Below are the detailed steps we have followed and plan to follow and shown in the Figure 1:

3.1 Data Collection

The workflow for our project involves several key steps aimed at preparing the data, building baseline models like (Naive Bayes, Random Forest, Decision Tree, Gradient Boost, and XGBoost) and evaluating the performance of a pre-trained language model named Bidirectional Encoder Representation from Transformers (BERT) [5]. Below are the detailed steps that are followed.

3.2 Exploratory Data Analysis (EDA)

Our initial step involved performing Exploratory Data Analysis (EDA) [3] on the CFPB dataset. For instance, we analyzed the frequency of complaints by category, discovering a higher volume of complaints in "Credit reporting" compared to other categories such as debt collection, mortgages and loans, retail_banking, and other_services as shown in Figure 2. This insight helped prioritize our preprocessing efforts to ensure diverse representation across categories.

3.3 Data Preprocessing

In the data pre-processing stage, several essential steps are executed to refine the dataset. Initially, missing data (Null values) and duplicates are handled to ensure data integrity. Then, class consolidation condenses the 21 product areas into five major classes, simplifying the classification task. To address class imbalance, a sampling technique is applied, enhancing model performance. Text data undergoes tokenization to break it into individual words, aiding model comprehension. Additionally, stopwords are removed to focus on meaningful words, refining textual analysis. These steps collectively prepare the dataset for effective analysis and modeling.

3.4 Dataset splitting

During dataset splitting, the preprocessed dataset is divided into training and testing sets using an 80-20 ratio. This means that 80% of the data is allocated for training the model, while the remaining 20% is reserved for evaluating its performance (Testing). This balanced partitioning ensures that the model is trained on a sufficient amount of data while maintaining an independent subset for unbiased testing. Such a systematic approach to dataset splitting is essential for developing and validating robust models effectively.

3.5 Baseline Model Building

For implementing the baseline models, we intend to utilize the scikit-learn library. Scikit-learn offers efficient implementations of various machine learning algorithms, including Multinomial Naive Bayes, Random Forest, Decision Tree, Gradient Boosting, and XGBoost. Leveraging scikit-learn's capabilities, we will train these models on the preprocessed dataset and assess their performance using metrics like accuracy, precision, recall, and F1-score. This standardized approach facilitates straightforward experimentation and comparison of different models, allowing for the identification of the most efficient model for further optimization and refinement.

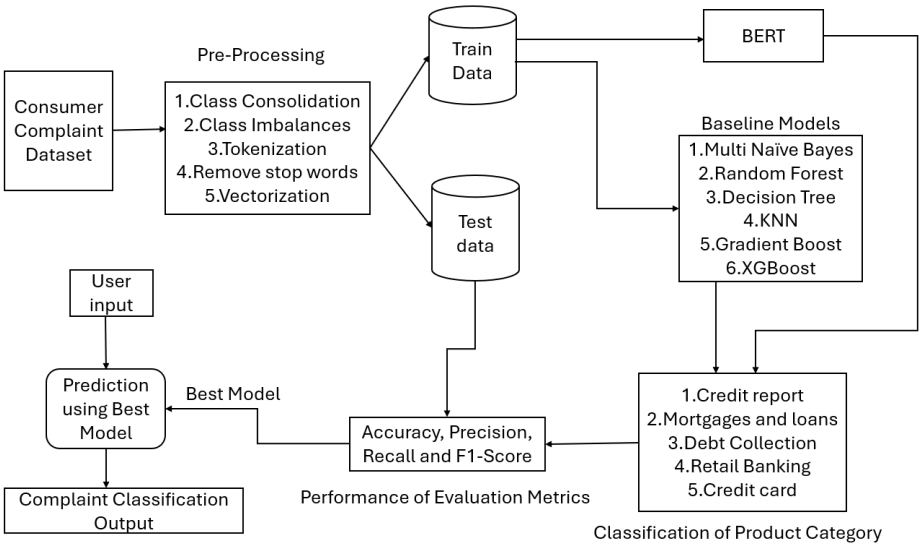

Figure 1: Architecture Diagram

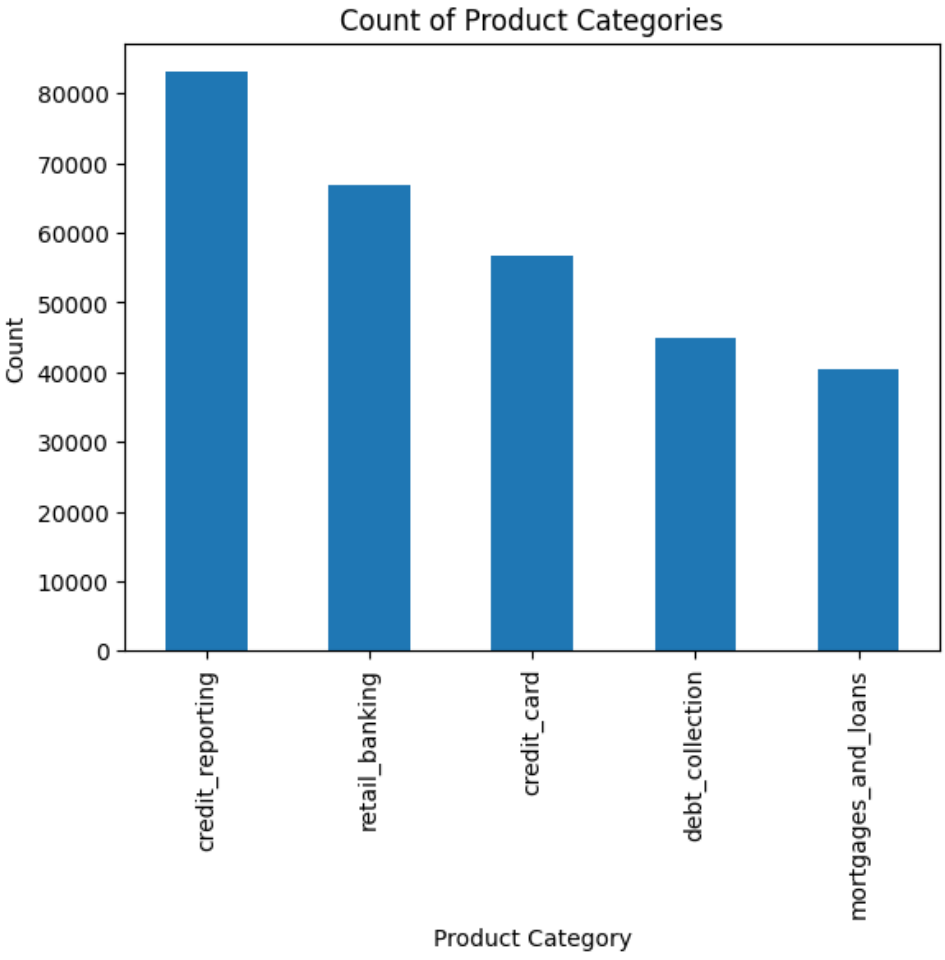

Figure 2: Frequency of Complaints based on Category

3.6 Pre-trained Language Model (PLM) Implementation

In implementing the Pre-trained Language Model (PLM), BERT [5] will be employed as the chosen model. Initially, BERT [5] undergoes fine-tuning using the training dataset to adapt to the specific task requirements. This fine-tuning process adjusts the model's parameters to better comprehend the dataset nuances. Subsequently, the performance of the fine-tuned BERT [5] model is assessed using the test dataset. This evaluation gauges the model's effectiveness in accurately processing and understanding the textual data, providing insights into its suitability for the given task.

3.7 Model Evaluation

For model evaluation, a fresh and unseen dataset will be employed to assess the performance of both the baseline models and the fine-tuned BERT model. By utilizing this new dataset, unbiased evaluation of the models' performance can be ensured. Comparative analysis of the performance metrics will then be conducted to discern the effectiveness of each approach. This evaluation process is crucial for determining which model, whether it be the baseline models or the fine-tuned BERT model, is better suited for the task at hand, thereby informing subsequent decision-making and model refinement efforts.

3.8 Final Implementation

In the final implementation stage, the model showcasing the highest performance based on evaluation metrics will be chosen for prediction on new consumer complaints. This selection process ensures the adoption of the most effective model for accurate complaint classification. By prioritizing the model with superior performance, we aim to enhance the reliability and precision of the classification system, thereby improving the overall handling of consumer complaints.

4 Data

For this project, we utilized a real-time dataset obtained from the Consumer Financial Bureau, an official website of the United States Government. The dataset spans a decade of consumer submissions, ranging from December 2011 to February 2024, comprises 4,858,723 rows and 18 columns, including features such as Date received, Product, Sub-product, Issue, Sub-issue, Consumer complaint narrative, Company public response, Company, State, ZIP code, Tags, Consumer consent provided?, Submitted via, Date sent to company, Company response to consumer, Timely response, Consumer disputed, and Complaint id. The relevant features that we are considering for the classification are Product, Sub-product and Consumer complaint narrative.

Data Preprocessing:

4.1 Cleaning and Filtering

Complaints lacking consumer narratives were removed from the dataset, focusing only on instances with detailed descriptions. Duplicate narratives were identified and eliminated.

4.2 Categorization

Initially, the dataset contained twenty-one distinct product categories. Through careful observation, these categories were consolidated into five main classes: Credit Reporting, Debt Collection, Mortgages and Loans, Credit Cards, and Retail Banking, simplifying the classification process.

4.3 Text Data Preprocessing

Stop words and irrelevant punctuation were removed from narrative texts to emphasize meaningful content. Lemmatization was applied to standardize words to their base or root form, reducing dimensionality and enhancing model performance.

4.4 Vectorization

Text data pertaining to product categories underwent vectorization, transforming words into numerical frequencies suitable for machine learning models.

4.5 Balancing Classes

Initially imbalanced, the data was balanced by randomly selecting data from each category to match the count of the category with the lowest representation. This step aimed to prevent model bias towards overrepresented classes.

4.6 Data Splitting

The refined dataset was split into training and testing sets to evaluate the performance of various machine learning models. Different data sizes were used during the training process, and data was filtered based on date parameters, such as from March 2020 to March 2023.

4.7 Potential Problems

Imbalanced data distribution may lead to biased model performance. The dataset's size may pose challenges in terms of computational resources and processing time.

5 Evaluation Plan

In our evaluation plan, we will compare the models simultaneously, analyzing their performance side by side based on key metrics. This approach allows us to directly assess which model best meets our criteria for accuracy and efficiency in classifying consumer complaints.

Metrics

5.1 Accuracy

This primary metric reflects the overall percentage of complaints correctly classified by the model. It provides a high-level view of the model's effectiveness in complaint categorization.

5.2 Precision (per class)

This metric delves deeper, measuring the proportion of complaints assigned to a specific category that truly belong there. It's crucial to avoid misdirected investigations.

5.3 Recall (per class)

Focusing on completeness, recall measures the percentage of relevant complaints within a category that the model correctly identifies. This ensures all significant complaints are captured.

5.4 F1-Score (per class)

This balanced metric combines precision and recall, offering a comprehensive view of performance for each complaint category. It's particularly valuable in datasets with class imbalance.

5.5 Confusion Matrix

This visual tool displays the model's classifications, highlighting correct and incorrect categorizations per category. It helps identify areas for improvement and potential patterns of confusion between specific complaint types.

6 Results

Several baseline models were trained and validated, including Multinomial Naive Bayes, Random Forest, Decision Tree, KNN, Gradient Boost, and XGBoost. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated for each model. The results are summarized in the Table 1 below:

Table 1: BERT and Base line Models Evaluation Metrics on Test Data

Model	Accuracy	Precision	Recall	F1-Score
Multinomial NB	0.82	0.79	0.78	0.78
Random Forest	0.85	0.86	0.79	0.83
Decision Tree	0.83	0.78	0.79	0.79
KNN	0.87	0.99	0.84	0.91
Gradient Boost	0.83	0.82	0.80	0.81
XGBoost	0.88	0.87	0.86	0.86
BERT	0.78	0.72	0.76	0.74

The XGBoost model exhibited superior performance across all metrics, achieving the highest scores in accuracy, precision, recall, and F1-score. Conversely, the Decision Tree model showed lower performance compared to other algorithms across these metrics. While Multinomial Naive Bayes, Random Forest, and Gradient Boosting displayed competitive results, they were slightly outperformed by XGBoost, especially in terms of accuracy and precision. Notably, KNN achieved exceptionally high precision, though it was slightly less accurate than XGBoost. These results suggest that XGBoost is the most suitable model for classifying consumer complaints, providing a balance of high accuracy and robust performance across all evaluated metrics.

Continuing our analysis, the validation and test accuracies for each model are presented in the Table 2 below:

Table 2: Validation and Test Accuracies for baseline models

Model	Validation Accuracy	Test Accuracy
Multinomial Naive Bayes	0.81	0.82
Random Forest	0.81	0.85
Decision Tree	0.75	0.83
KNN	0.34	0.87
Gradient Boost	0.82	0.83
XGBoost	0.85	0.88

XGBoost and Multinomial Naive Bayes show robust validation accuracies, with XGBoost leading at 0.85 and maintaining its lead in the test accuracy with a high score of 0.88. Random Forest, though matching the validation accuracy of Multinomial Naive Bayes, slightly improves upon moving to the test set with an accuracy of 0.84. This indicates its good generalization from validation to test data.

Conversely, KNN, despite achieving a high test accuracy of 0.87, shows a markedly low validation accuracy of 0.34, suggesting issues with overfit-

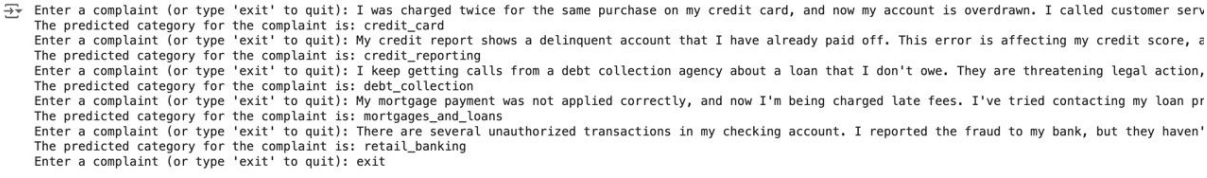

Figure 3: Sample Prediction Using XGBoost

ting or inconsistencies in the model's generalization capabilities. The Decision Tree, while having the lowest validation accuracy at 0.75, surprisingly shows a higher test accuracy at 0.83, which may indicate its potential for better adaptation to unseen data.

These findings underscore XGBoost's effectiveness in consistently delivering high performance across both validation and test datasets, affirming its suitability for the task of classifying consumer complaints in the financial sector.

In addition to the baseline models, we implemented the BERT (Bidirectional Encoder Representations from Transformers) model for text classification. BERT's advanced capabilities in understanding context and nuance in text make it a powerful tool for this task. Our fine-tuned BERT model achieved an F1 score of 0.77 on the test dataset, demonstrating its competitive performance.

The confusion matrix for BERT revealed varying levels of performance across different complaint categories, with particularly strong results in 'Credit Reporting' and 'Retail Banking'. The detailed classification report highlighted BERT's precision, recall, and F1-score for each class, with an overall accuracy of 0.775. These results indicate that while BERT excels in certain categories, there are areas for improvement, particularly in 'Debt Collection', where it displayed weaker performance.

These outcomes underscore the potential of advanced transformer models like BERT in handling complex text classification tasks. However, they also highlight the necessity for targeted adjustments to model training and data preprocessing to enhance accuracy and efficiency. The mixed results suggest that while BERT is powerful, its application requires careful tuning and validation to maximize effectiveness in specific contexts such as consumer complaint classification. Future work should focus on optimizing hyperparameters and exploring additional data preprocessing techniques to further leverage BERT's capabilities.

To complement BERT, we also implemented a real-time prediction system using XGBoost, allowing users to input complaints and receive predicted categories immediately. This practical application of our model demonstrates its usability and effectiveness in categorizing consumer complaints. Figure 3 shows sample predictions from this system, highlighting its accuracy.

In conclusion, our project demonstrates the feasibility and effectiveness of using advanced machine learning models for automating the classification of consumer complaints. Both BERT and XGBoost have shown strong potential, with XGBoost providing the most consistent results. Future work will focus on refining these models, exploring ensemble methods, and expanding the system's capabilities to handle a broader range of complaint types.

By leveraging these advanced techniques, we aim to significantly improve the efficiency of complaint resolution processes in the financial sector, leading to quicker resolutions and enhanced customer satisfaction.

7 Discussion

This project provided key insights and exposed several challenges that could guide future developments in consumer complaint classification.

Challenges and Lessons Learned

A major challenge was the imbalanced data which impacted the generalizability of the models. Advanced techniques such as synthetic data generation (SMOTE) [6] could be explored in future work to improve data balance. Additionally, fine-tuning the BERT model proved resource-intensive; optimizing this process is essential for practical applications.

Recommendations for Future Work

If starting this project anew, an increased focus on initial exploratory data analysis (EDA)[7] would be beneficial. This could provide deeper insights for more informed data preprocessing and feature

engineering decisions. Additionally, exploring a broader range of machine learning models early in the project could help identify optimal solutions sooner.

For those continuing this work, the following steps are recommended:

1. Enhanced Model Exploration: Investigate newer models and advanced ensemble techniques to improve accuracy and robustness.

2. Hyperparameter Optimization: Employ automated tools like Hyperopt for more efficient tuning, particularly for complex models like BERT and XGBoost.

8 Project Code Link

Click here to access the code for Consumer Complaint Classification

References

1. Thomas, N. T. (2018, September). A LSTM based Tool for Consumer Complaint Classification. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2349-2351). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8554857&isnumber=8554361

2. González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012. https://doi.org/10.48550/arXiv.2005.13012

3. Naik, P. K., T, P., S, C., S, J., & Balan, S. (2023a). Consumer Complaints Classification Using Machine Learning & Deep Learning. International Research Journal on Advanced Science Hub, 5(Issue 05S), 116–122. https://doi.org/10.47392/irjash.2023.s015

4. Bozyigit, F., Dogan, O., & Kilinc, D. (2022). Categorization of customer complaints in food industry using machine learning approaches. Journal of Intelligent Systems: Theory and Applications, 5(1), 85–91. https://doi.org/10.38016/jista.954098

5. Hugging Face. (n.d.). Transformers documentation: BERT. Retrieved from https://aclanthology.org/N19-1423/

6. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://www.jair.org/index.php/jair/article/view/10302

7. Velleman, P. F., Hoaglin, D. C. (2012). Exploratory data analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 3. Data analysis and research publication (pp. 51–70). American Psychological Association https://psycnet.apa.org/record/2011-23865-003