ConsumerReports

Natural Language Processing

Davide Teixeira

Eduardo Oliveira

Máximo Pereira

Dataset

The dataset consists of customer complaints submitted to financial institutions. It includes various fields such as:

- **Unnamed: 0**: Index column.
- **product_5**: category related to the complaint.
- **narrative**: detailed description of the complaint
- **Product**: specific type of product/service involved.
- **Date received**: The date the complaint was registered.
- **Sub-product**: A more granular classification of the product.
- **Issue**: The main problem described in the complaint.
- **Sub-issue**: A more detailed classification of the issue.
- **Company**: The entity that received the complaint.
- **State**: The state where the complaint originated.
- **Timely response?**: Indicates whether the company responded within an appropriate timeframe.

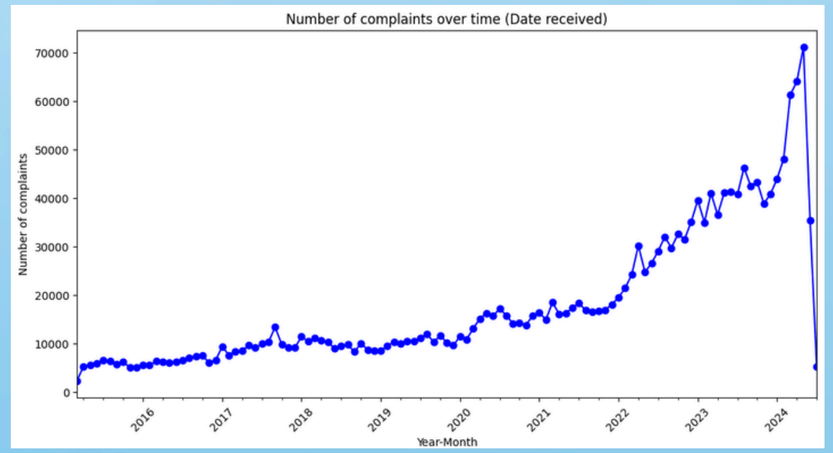
Exploratory Data Analysis

Dataset size

- The dataset contains 2,023,066 consumer complaints related to financial services.

Trends over time

- The number of complaints fluctuates, with peaks in 2024.

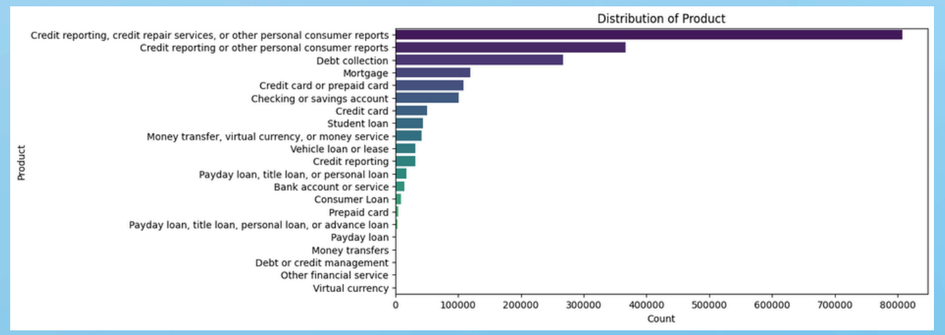


Product Distribution

- Most complaints are related to Credit reporting, credit repair services, or other personal consumer reports, representing a significant portion of the total.
- Other major complaint categories include Debt collection, Mortgage, and Credit card services.

Missing values

- Some columns have missing values, such as Sub-product (52,206 missing), Sub-issue (230,559 missing), and State (7,344 missing).



Exploratory Data Analysis

Text Analysis

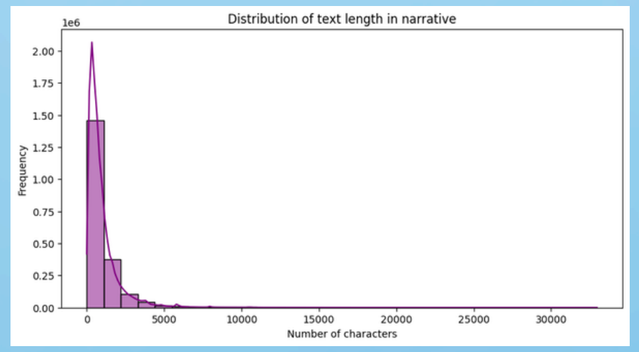
- The analysis of complaint narratives reveals that words like “credit report” and “credit reporting” appear frequently, indicating common consumer concern.

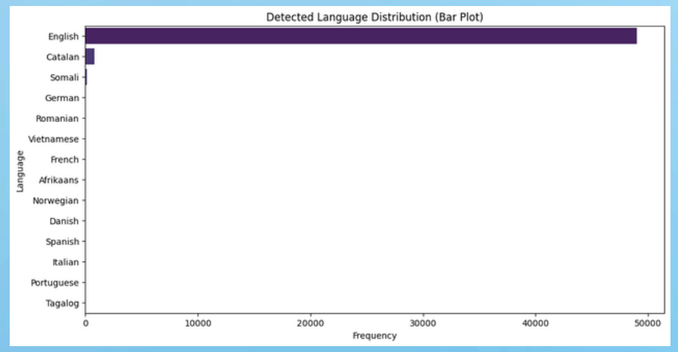
Text length distribution

- The Issue, Sub-issue, and Narrative fields have different text length distributions, with the Narrative field being significantly longer than the others.

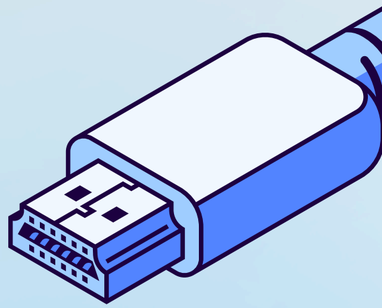
Language detection

- The majority of complaints are written in English, with a small number in Catalan and Somali (and a few others). However, since these languages appear in very low frequencies, they can be considered outliers and mitigated accordingly.





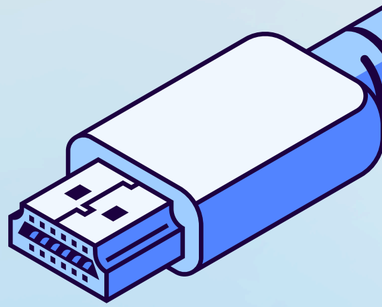
Pre-processing



- Enumerate common text variables
- Drop null columns and rows
- Convert text to lowercase

Target Variable

- Credit Reporting → 0
- Debt Collection → 1
- Loans → 2
- Bank Accounts and Services → 3
- Credit Card Services → 4



Pre-processing

- Remove Special Characters using Regular expressions
- Remove extra spaces
- Remove stopwords
- Lemmatize → English Words

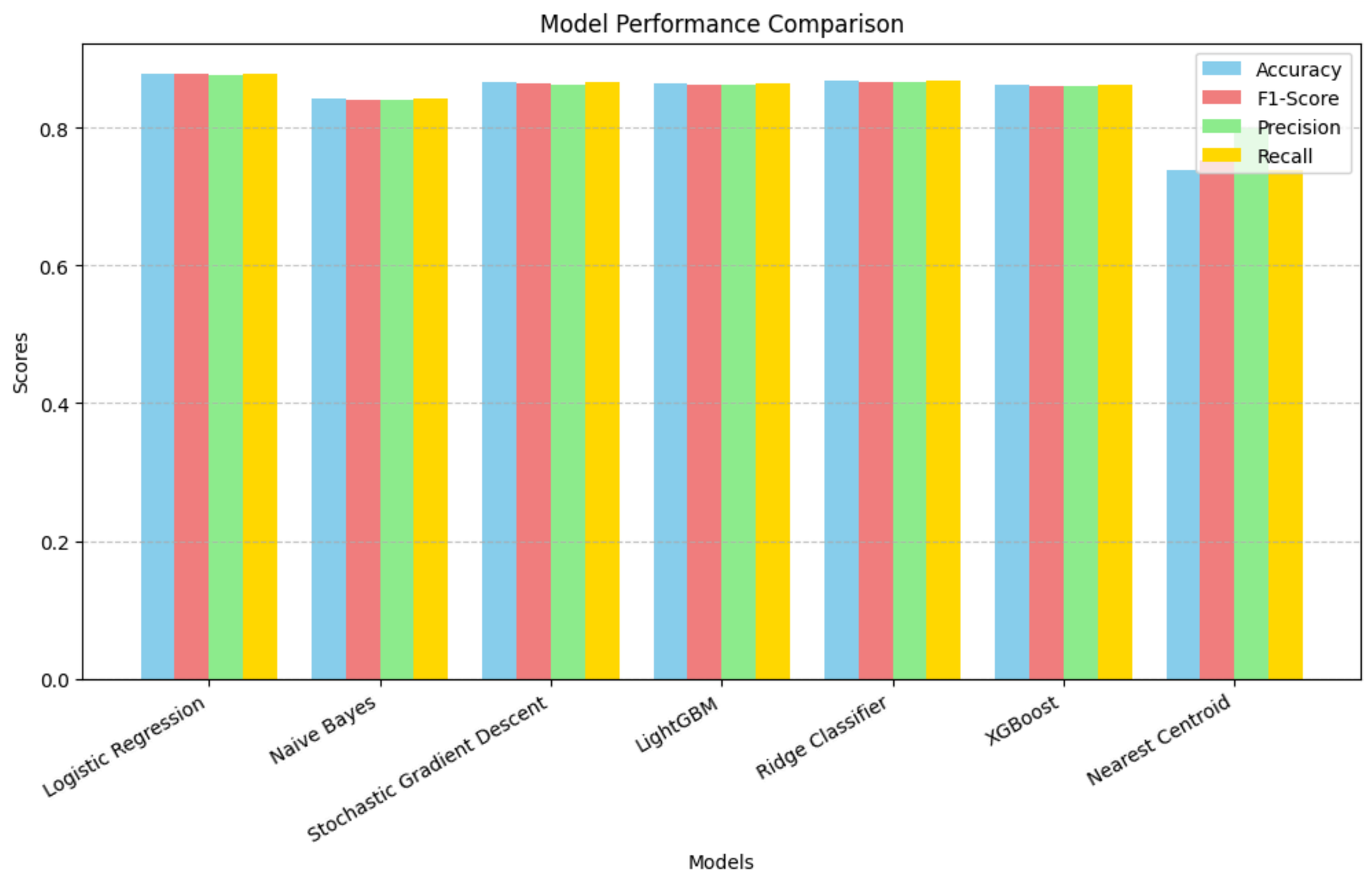
Regular Expression

[^a-zA-Z0-9\s]

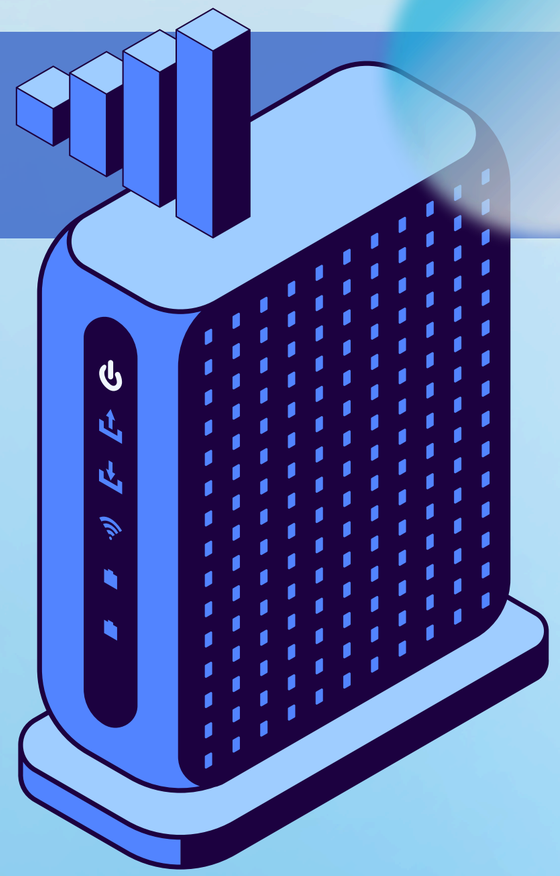
Experimental Results

- Randomized Train-Test
  - split of 80%-20%
- Conversion of text to TF-IDF features
  - TfidfVectorizer from sklearn
- Exploration of several models, with the final choices being:
  - Logistic Regression
  - Naïve Bayes
  - Stochastic Gradient Descent
  - LightGBM
  - Ridge Classifier
  - XGBoost
  - Nearest Centroid

Experimental Results



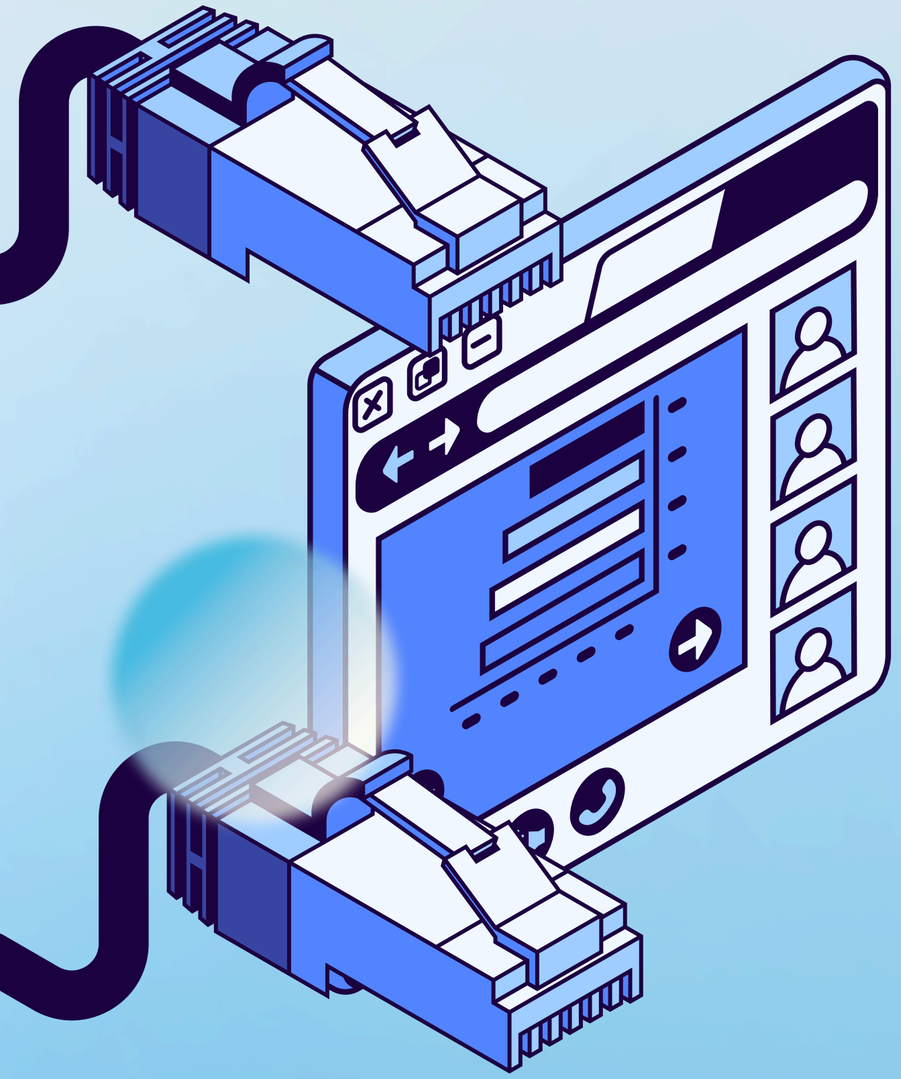
Error Analysis



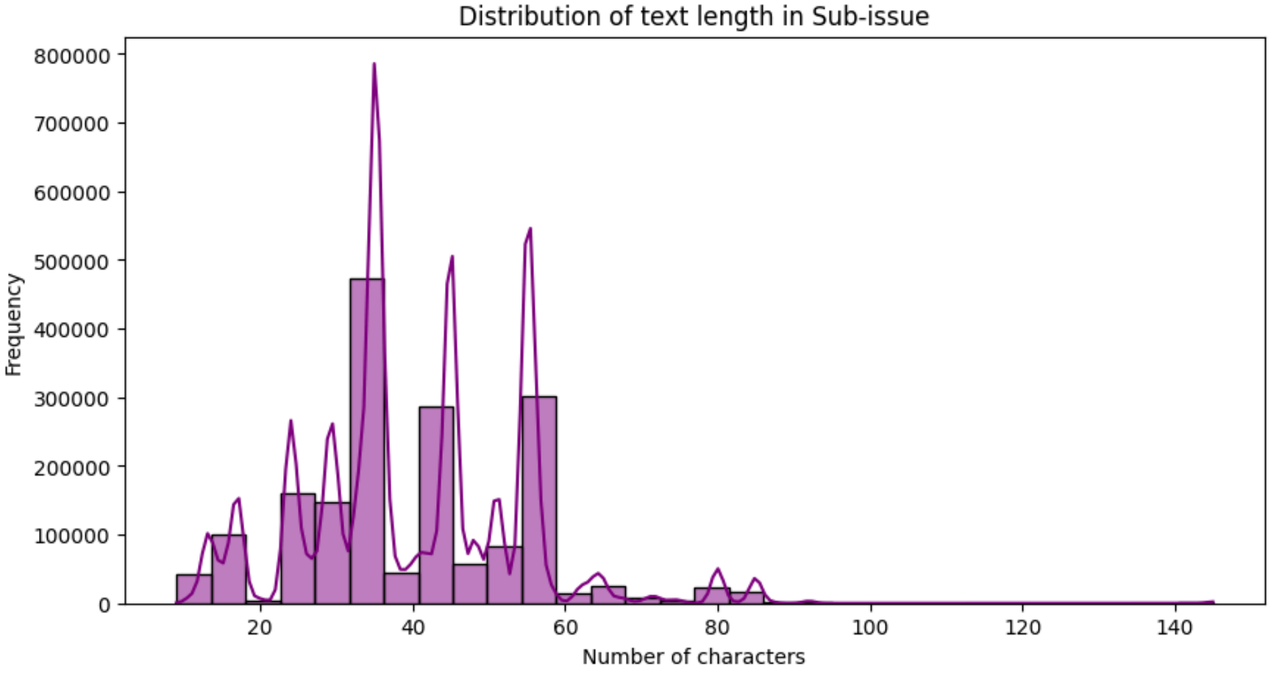
Performance of the Models

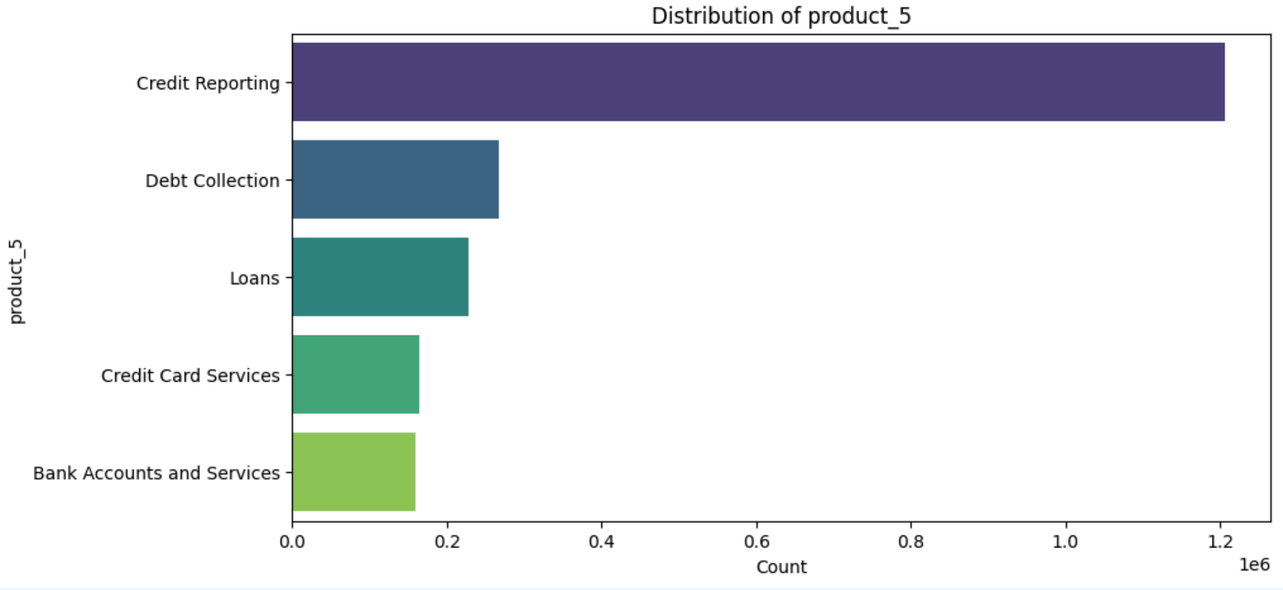
- Logistic Regression (88%) & Ridge Classifier (87%)
  - modelled linear decision boundaries effectively for text classification
- Stochastic Gradient Descent (86%)
  - used due to efficiency in large-scale datasets and it worked well optimizing a hinge loss
- LightGBM (86%) & XGBoost (86%)
  - learn non-linear relationships and handle sparse TF-IDF data effectively
- Naïve bayes (84%)
  - used as a baseline, assumes word independence
- Nearest Centroid (80%)
  - poor performance due to being too simplistic for high dimensional data
  - it works well when categories are well separated and have clear differences, which usually isn't the case with text

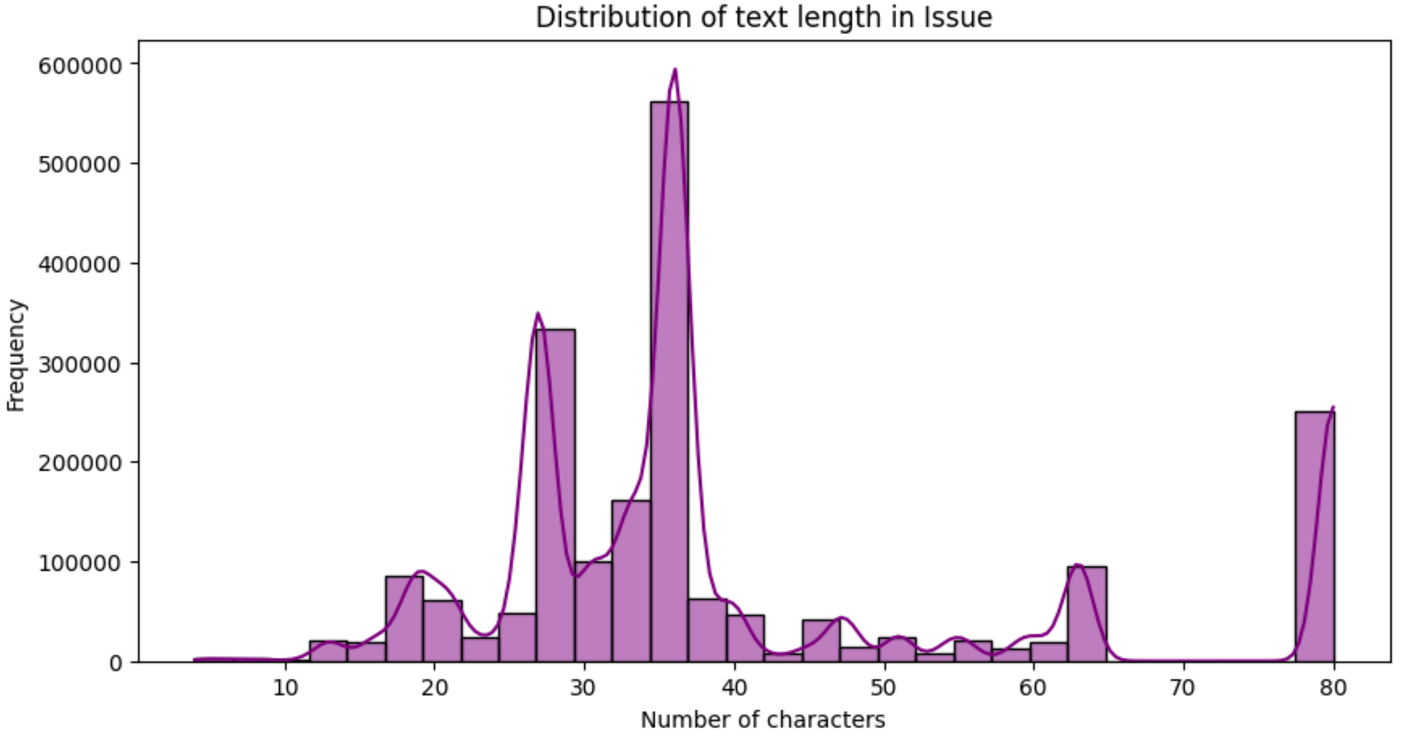
Conclusions



- Several techniques were utilised across the development of the project
- Some algorithms were explored, but weren't used due to taking too long to train:
  - Random Forest
  - Neural Networks
  - K-Nearest Neighbours
  - Support Vector Machine
- Results were moderate, not great across every model, some possible answers as to why:
  - lack of hyperparameters optimizing
  - relatively small data sample size due to the time required for training
  - lack of exploration of more complex algorithms, due to time complexity








```
Category distribution:
Product
Credit reporting, credit repair services, or other personal consumer reports    807291
Credit reporting or other personal consumer reports                             366397
Debt collection                                                                 266842
Mortgage                                                                        119116
Credit card or prepaid card                                                     108669
Checking or savings account                                                     100447
Credit card                                                                      50372
Student loan                                                                     44241
Money transfer, virtual currency, or money service                               41503
Vehicle loan or lease                                                            32077
Credit reporting                                                                 31587
Payday loan, title loan, or personal loan                                        17238
Bank account or service                                                          14885
Consumer Loan                                                                     9461
Prepaid card                                                                      4669
Payday loan, title loan, personal loan, or advance loan                           3816
Payday loan                                                                       1746
Money transfers                                                                   1497
Debt or credit management                                                          904
Other financial service                                                            292
Virtual currency                                                                    16
```