Classifying Consumer Complaints Using Natural Language Processing

Capstone Project for Flatiron's School Data Science Bootcamp

by HENRY ALPERT

May 2021

FLATIRON SCHOOL

PROJECT OVERVIEW

About the Data

- Submitted by consumers to the **Consumer Financial Protection Bureau**, a "U.S. government agency that makes sure banks, lenders, and other financial companies treat you fairly."

- Consumers can submit a **narrative** of their complaint and are prompted to classify their complaint in four categories:
  - product
  - sub-product
  - issue
  - sub-issue

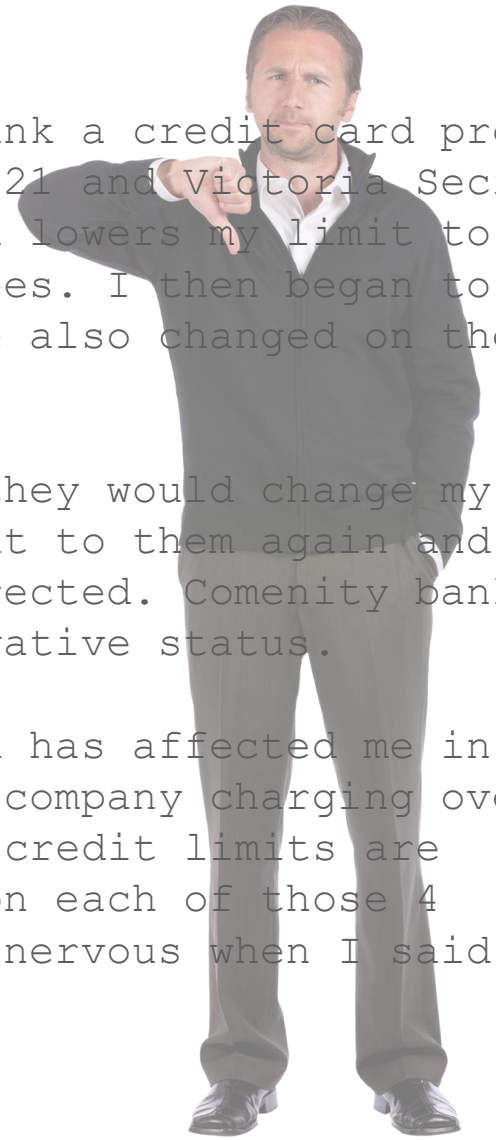
cfpb Consumer Financial Protection Bureau

Sample Narrative

Hello my name is XXXX XXXX, I am being scammed by Comenity bank a credit card provider for companies The Children's place, New York & Co. , Forever 21 and Victoria Secret. My original credit from XXXX was {$500.00} Comenity bank then lowers my limit to {$300.00} and began to charge overage fees along with late fees. I then began to pay close attention to my other cards to find that my limits were also changed on them as well incurring overages and late fees.

I reached out to the company Comenity bank they stated that they would change my credit limit to its original limits but did not. I reached out to them again and told them I will not summit any payment until my accounts are corrected. Comenity bank credit cards has impacted my credit scores plummeted to a negative status.

I'm currently paying the price due to the corruption in which has affected me in detrimental way. I am now in debt over {$2000.00} due to the company charging overage fees as well as late fees even through COVID-19. The initial credit limits are fluctuating tremendously and the company charges major fees on each of those 4 accounts. They are not willing to correct my account and was nervous when I said I had an attorney, that is the reason I'm reaching out to you.



Purpose of Project

- Develop a Natural Language Processing model which can use the narratives' text alone to categorize the complaints.

Business Case

- An NLP model will make the classification of complaints and their routing to the appropriate teams more efficient than manually tagged complaints.

EXPLORATORY DATA ANALYSIS

Data Overview

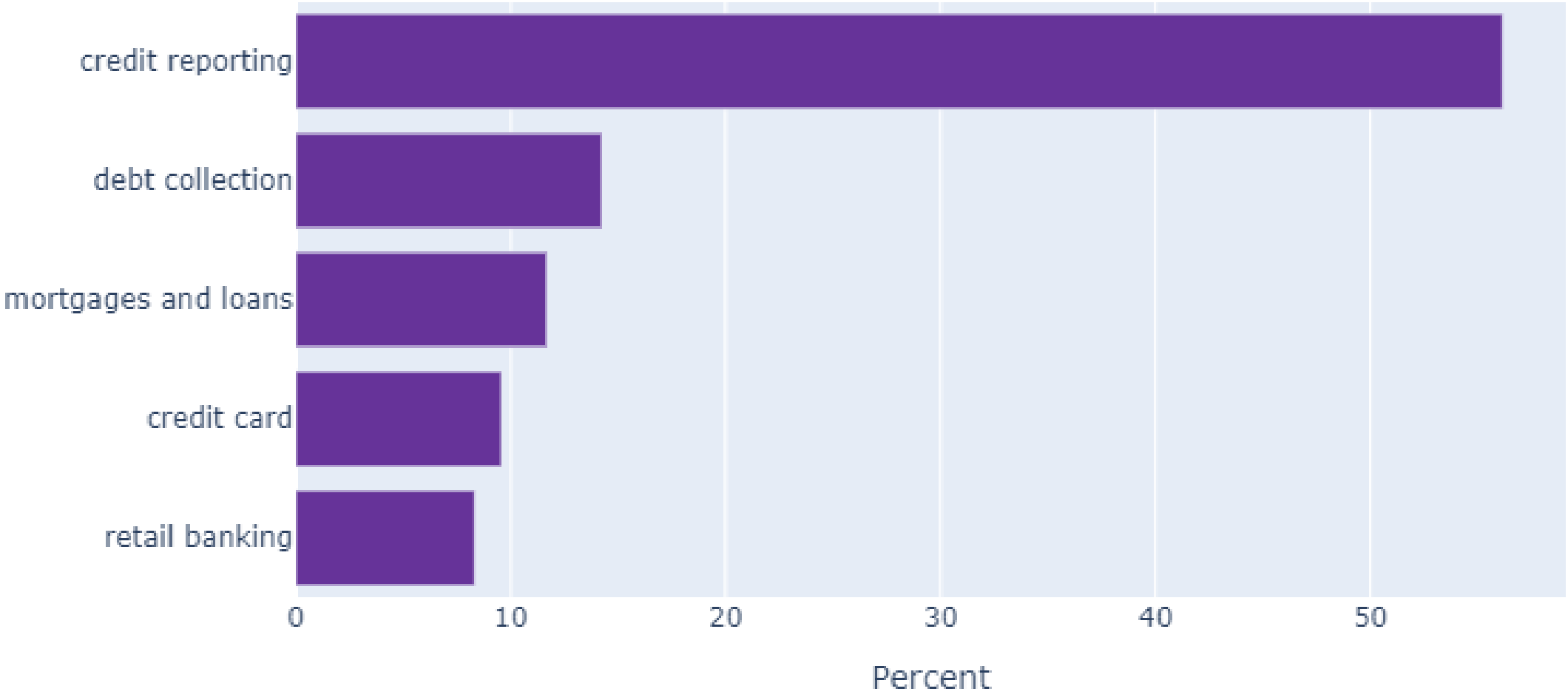
- Includes a year's worth of submissions from March 2020 to March 2021
- After removing submissions without narratives, about 162,400 complaints remained.
- Complaints were tagged with one of nine **product** areas.
- For the other categories (**sub-product**, **issue**, **sub-issue**), there were too many tags, often with too few instances, to be useful to train an NLP model.

Class Consolidation

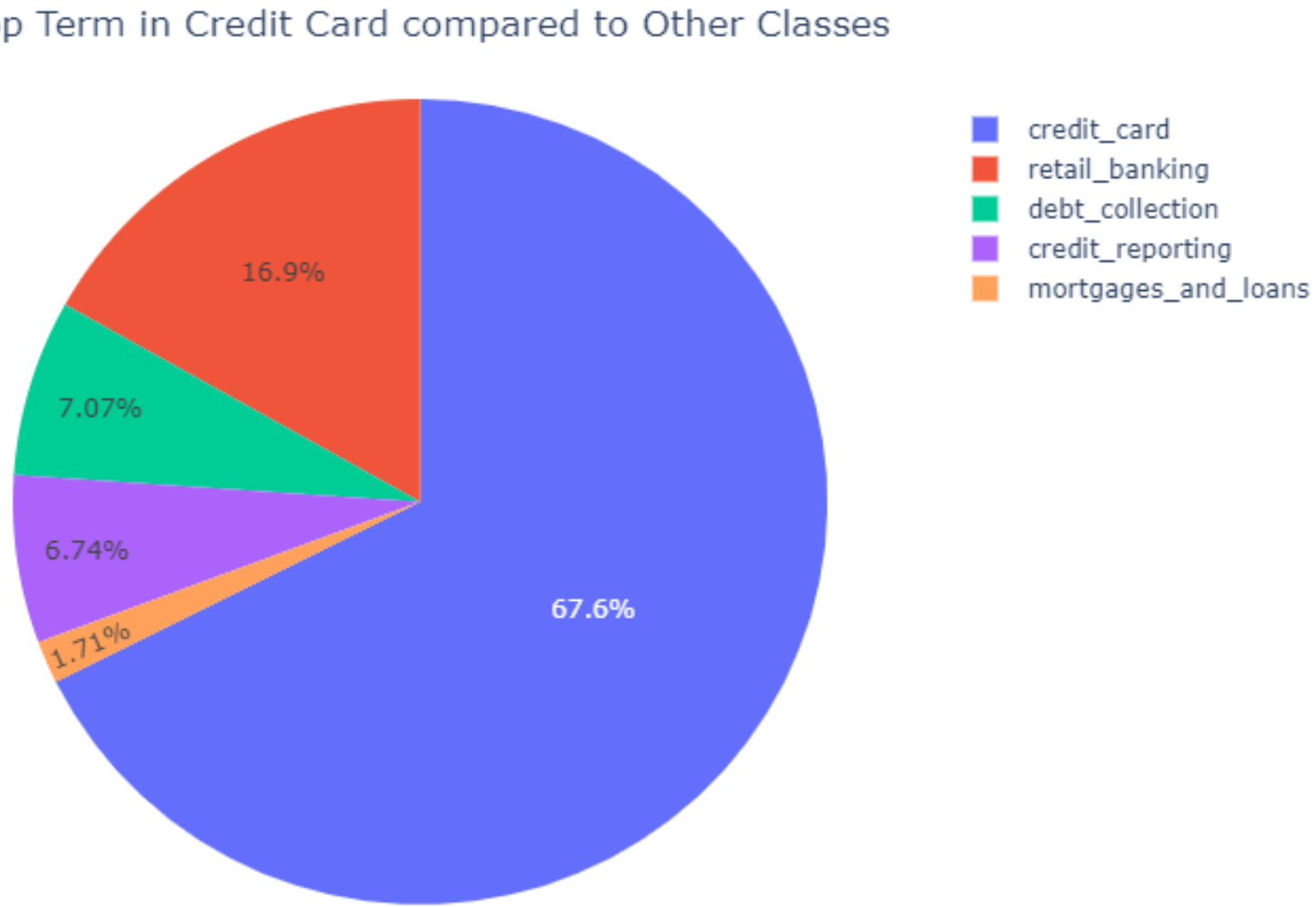
Consolidated the products into **five classes:**

1. credit reporting
2. debt collection
3. mortgages and loans
4. credit cards
5. retail banking

Class Imbalances



WORD PROMINENCE



MODELING

Modeling Process

Data Preparation

- Removed stopwords like “the” and “if”
- Lemmatized words (“banks” → “bank”)
- Vectorized data (transformed words into their numerical frequencies)
- Separated data into training set to train model and testing set to verify performance

Baseline Modeling

- Ran six different baseline models (Multinomial Naïve Bayes, Random Forest, Decision Tree, KNN, Gradient Boosting, XG Boost)

Modeling Process (cont'd)

Scoring

- Used **recall** as the primary metric since the five classes are imbalanced
- Looked for similar results between training and test results to minimize overfitting

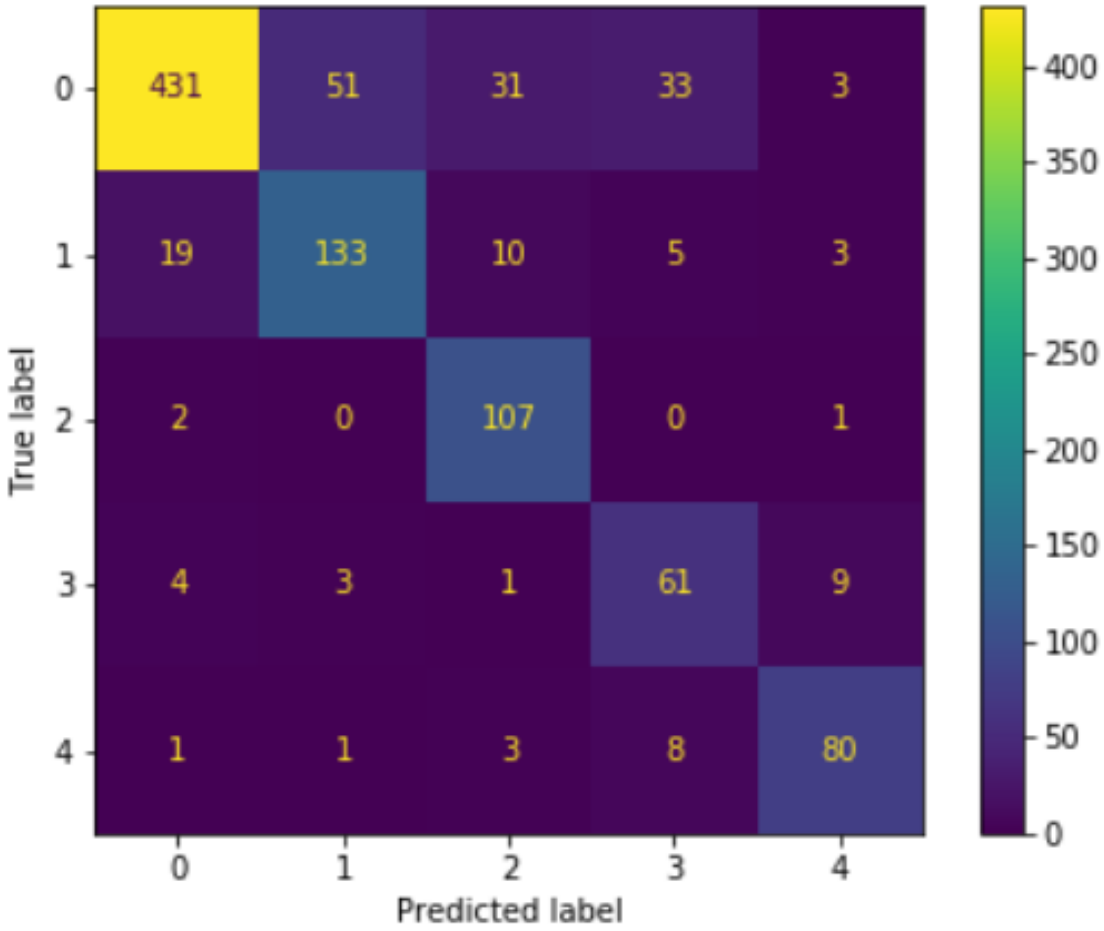
Refinement

- Experimented with various parameters on three of the best baseline models
- Multinomial Naïve Bayes had best overall score, classifying narratives correctly 86% of the time.

Running Model on Fresh Data

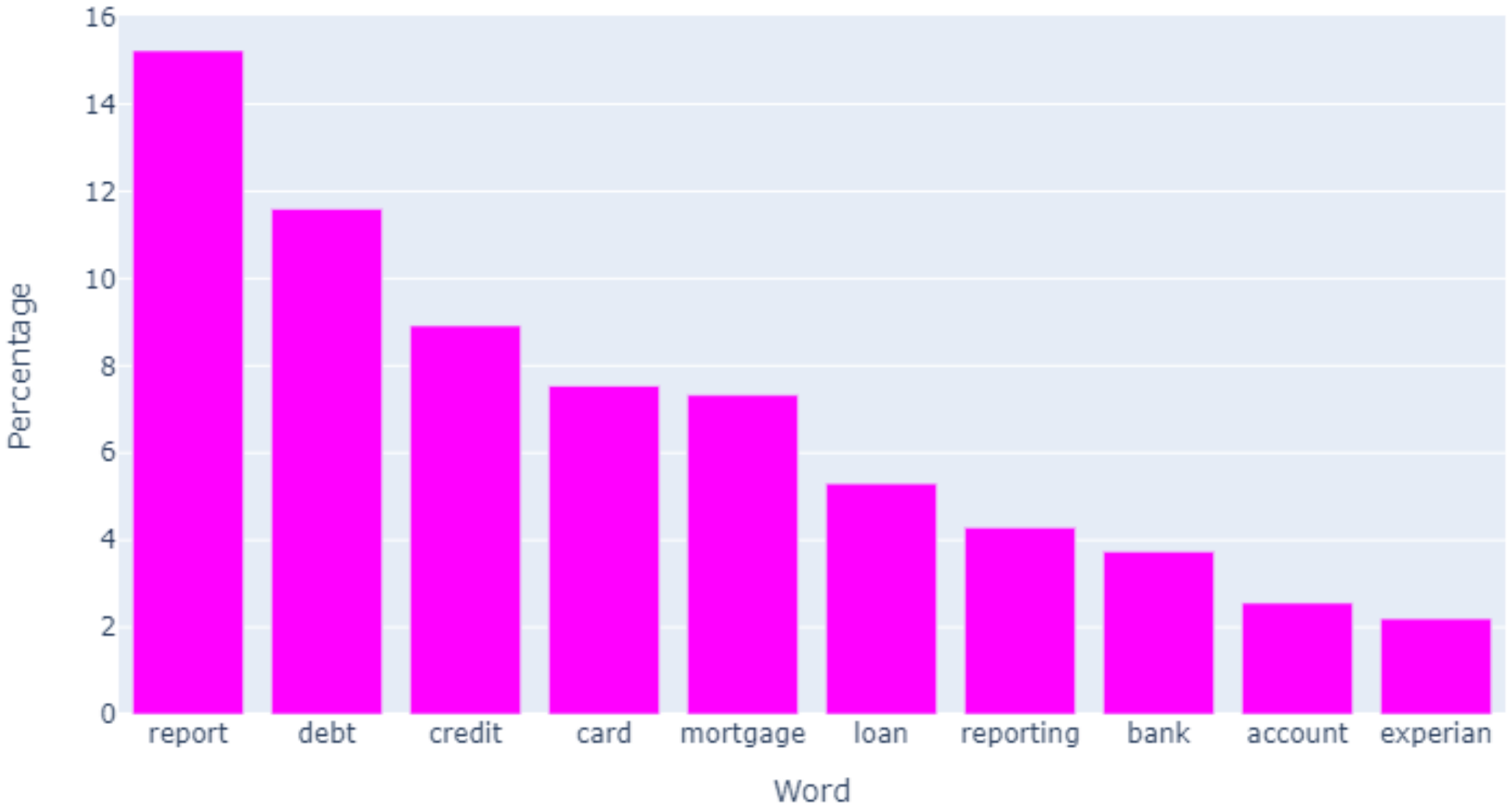
- Downloaded 1,000 new narratives from CFPB's API

- The trained model categorized class 2 narratives (mortgages and loans) particularly well – more than 97% correct.



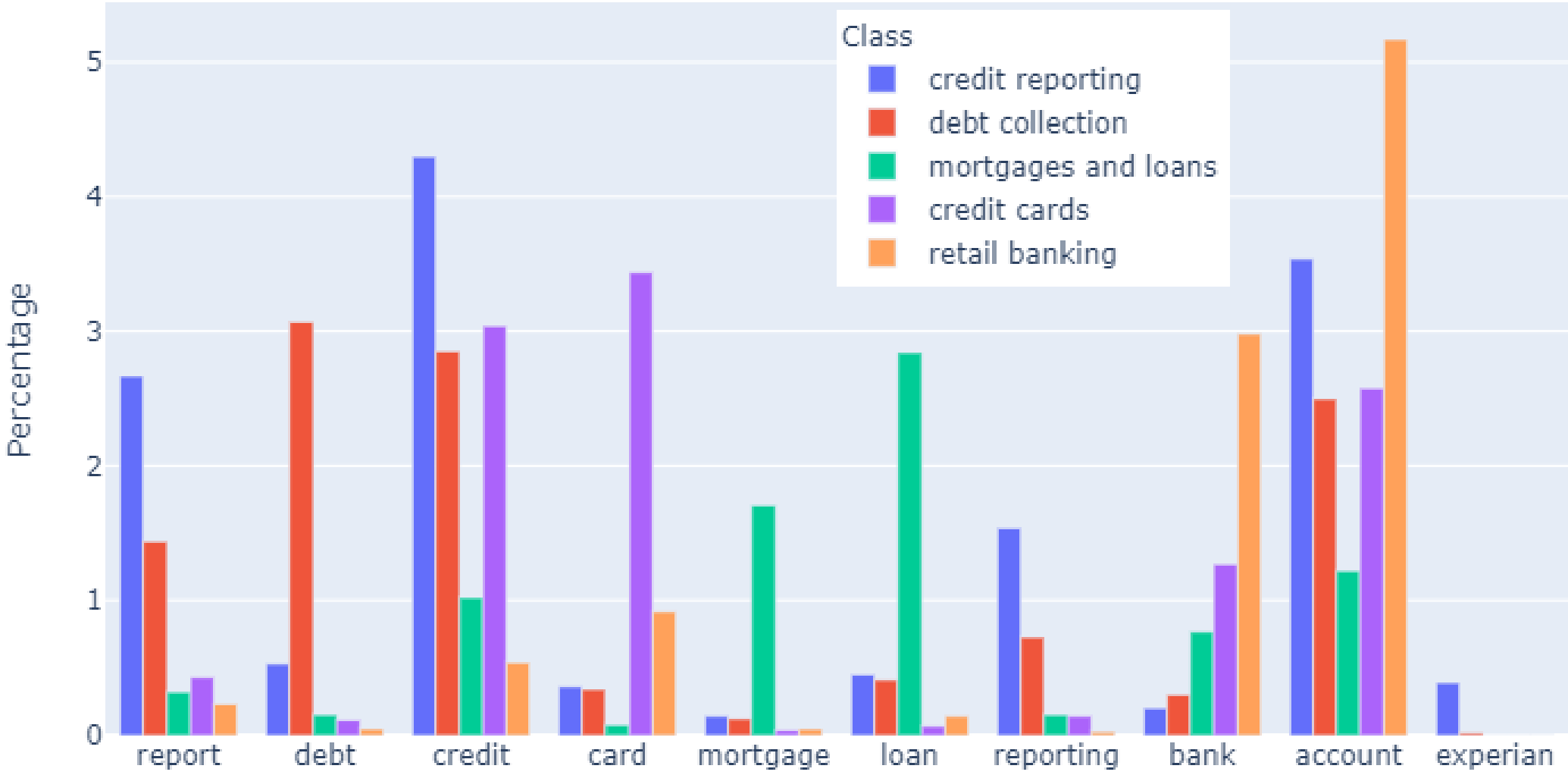
POST-MODELING EDA

Top 10 Most Important Features



Feature Importances:

Percentage of Word Prominence per Class



NEXT STEPS

Improve Business Case

- Since consumers classified their own complaints, ask CFPB employees to double-check narratives' classes, particularly those that the model misclassified
- Understand how the CFPB routes and processes consumer complaints and develop further modeling capabilities for **sub-product**, **issue**, and **sub-issue**

Refine Models

- Use more than one year's worth of data and further refine parameters
- Create Latent Dirichlet Allocation (LDA) model to develop new classification categories and learn if they might be useful to CFPB

CONTACT

Henry Alpert
halpert3@gmail.com

LinkedIn:
linkedin.com/in/henryalpert

GitHub Project Repo:
github.com/halpert3/nlp-classification-project

DATA SOURCE

Consumer Financial Protection Bureau - Consumer Complaint Database

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