Using Sentiment Analysis to Differentiate Student and Vehicle Loans: An Analysis in R

Motivation and Goal Setting

We were motivated to investigate the ties between complaint sentiment and loan type for three reasons. Firstly, loans are a relatable topic to most college students as they are a common option for students to afford education. Secondly, loans, like any large financial decision, can be stressful, and dealing with long-term loans is a task that can affect a person's emotions. Thirdly, the dataset we found from the Consumer Financial Protection Bureau (CFPB) was well-structured and organized in an easy-to-understand manner. We preferred to continue this project after confirming the data would be plausible to manipulate, clean, and interpret for our goal.

Our goal was not fully clear in the beginning. We knew we wanted to test sentiment in the use of different product types, but not necessarily specific loan types. We figured that by working backward, we would see which analysis would make the most sense to us. We came to the realization that we wanted to focus on student and vehicle loans as they represented an adequate amount of observations to perform a hypothesis test analysis, and they were the most common type of product loans faced by our student class.

After discussion, we came up with two research questions: First, do customers have different proportions of negative words in complaints about two types of products: "Student loan" and "Vehicle loan or lease"? Second, can we differentiate "Student loan" and "Vehicle loan or lease" complaints using classification models? We decided to use different statistical methods to solve these two questions, and the details will be demonstrated in the following paragraphs.

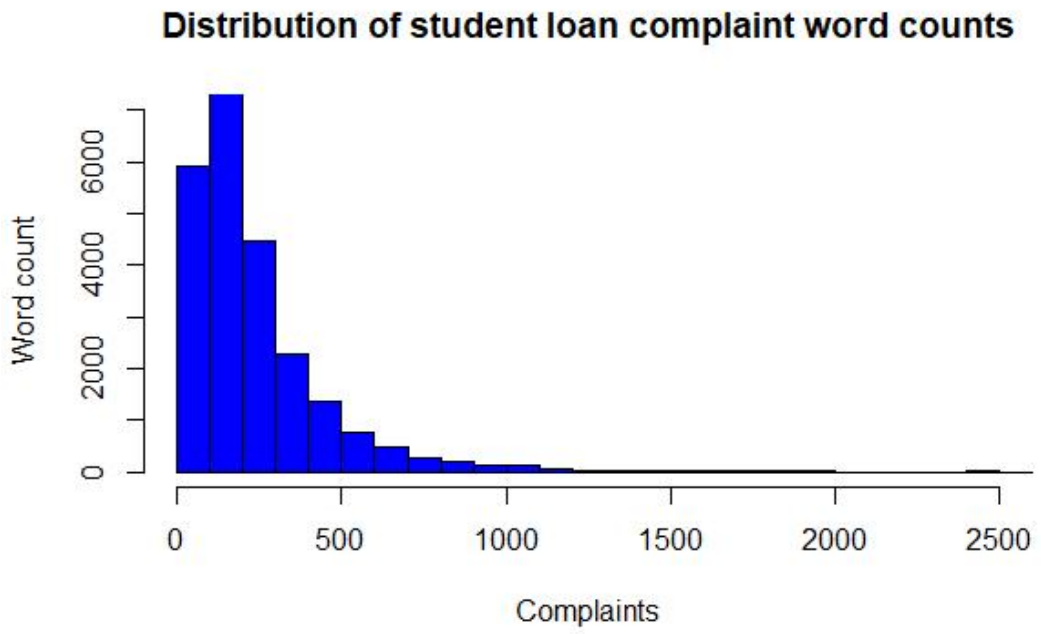
Data Collection and Cleaning

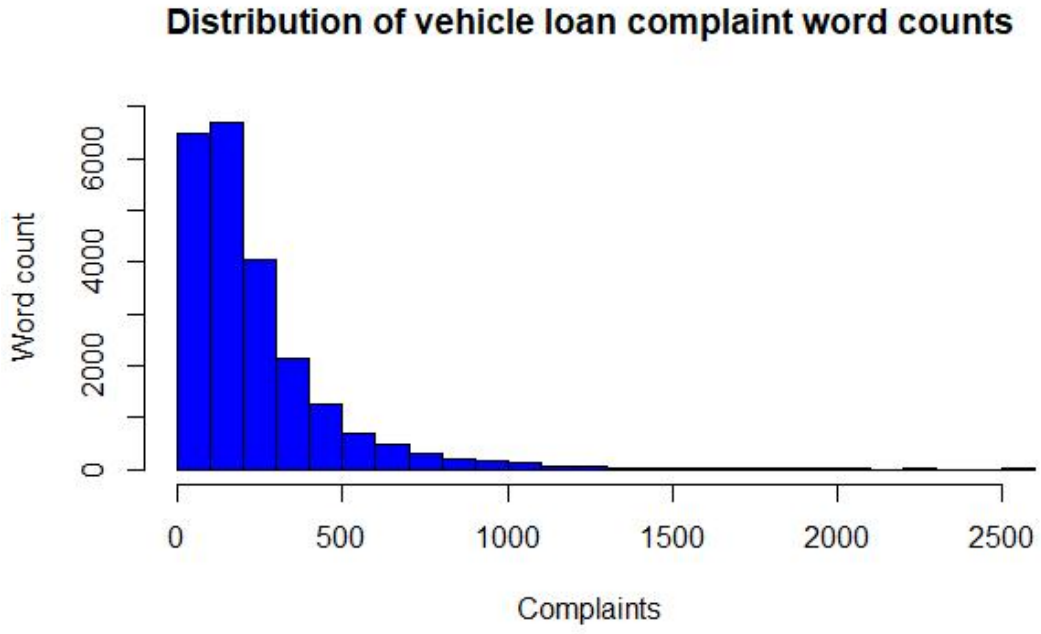
Initially, we tried to download the entire available dataset, but we soon found that it was too large to work with, especially when performing complex operations with the narrative portions. Therefore, we decided to filter the data to only include the "Student loan" and "Vehicle loan or lease" product categories since they are very relatable to us and contained a similar number of complaints. We also filtered out data that were collected before April 24, 2017, because that was the date of a major change in the way product and sub-product were categorized, as well as a few other data collection and reporting changes.

The data required minimal cleaning; there were few missing values, dates were consistently formatted, and most variables were collected via closed-ended questions, so there was no room for misspellings. The narrative complaints were already anonymized, using "X" in place of redacted information. We replaced redacted information with "RedactedDate" or "RedactedOther" depending on its format.

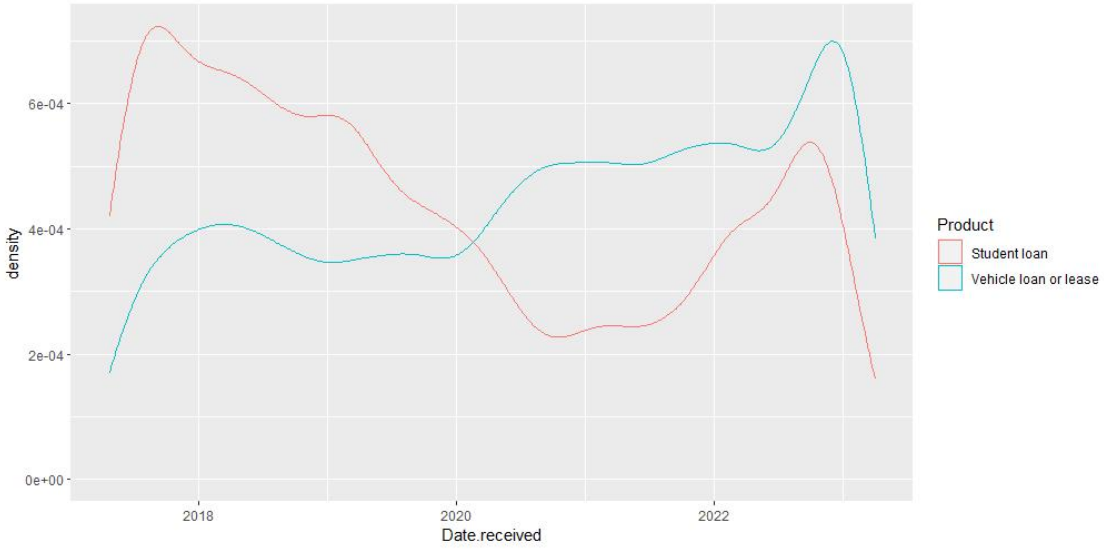
Exploratory Data Analysis

Once we had cleaned the data, we performed some exploratory data analysis to better understand its trends. The first aspect we explored was the length of the complaints submitted for the vehicle and student loans, respectively. Both categories have very similar means and quartile measurements, but the ranges differ slightly (max 6,327 versus 5,412 words, respectively). The standard deviations differed slightly, being approximately 247.6 and 278.97, respectively. The similarity in the distributions of the data led us to believe that our t-test test would reveal a slight difference, if any.





Here are the densities of both types of complaints over time:



Quantifying the Narratives

To compare the differences in the proportions of negative words in complaints about “Student loan” and “Vehicle loan or lease,” we need to have the proportion data first. However, the raw data do not include this variable, so we had to calculate it ourselves. We defined the proportion for each narrative to be the function:

$$\frac{the\ number\ of\ negative\ words}{the\ number\ of\ non-stop\ words} * 100\%$$

Notice we did not choose the total number of words in each narrative as the denominator because stopwords would dilute our measurement of negativity. Also, choosing the non-stopwords as the denominator would make the proportions generally higher and the results more evident.

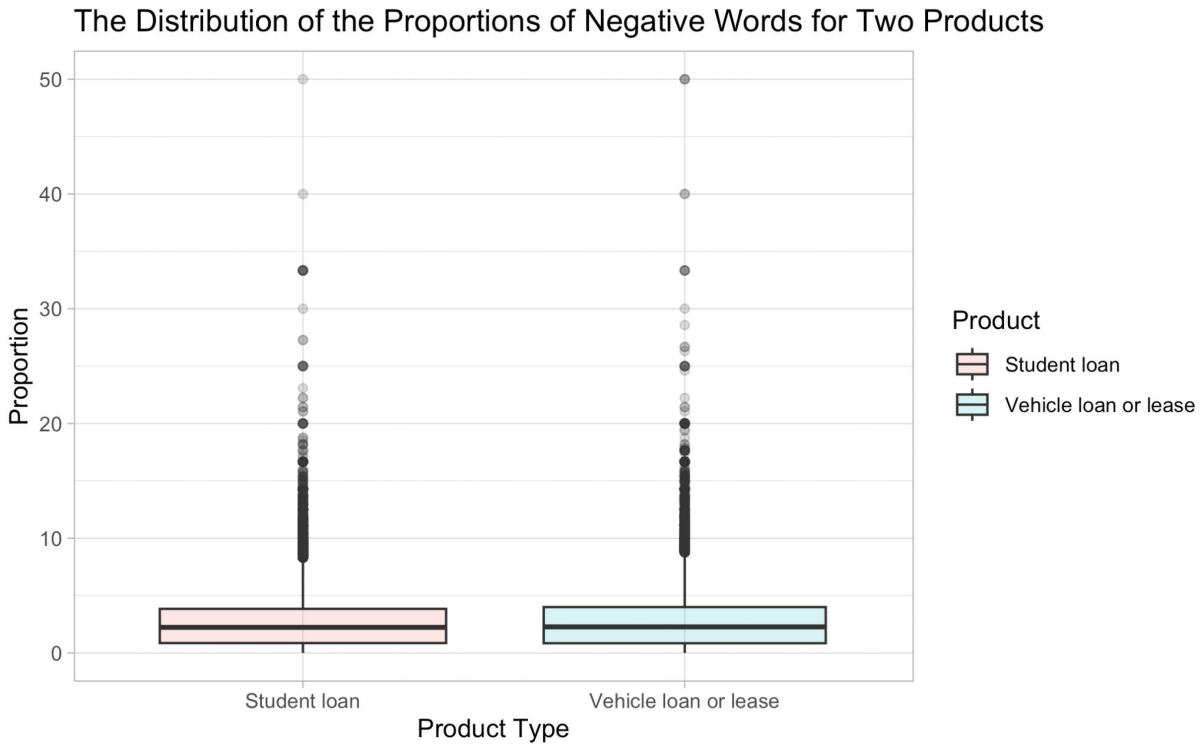
To count the number of negative words in each narrative, we first used Negative Word Dictionary found on GitHub [1][2]. The following are some examples of the words on the list:

```
"2-faced"       "2-faces"       "abnormal"
"abolish"       "abominable"    "abominably"
"abominate"     "abomination"   "abort"
```

Then we stemmed the narratives using the wordStem() function in the SnowballC package to match narratives and the Negative Word Dictionary better. By comparing whether the stemmed words in the narratives are in the Negative Word Dictionary (if yes, we count 1; if no, we count 0), we had the number of negative words in each narrative.

Finally, we applied the English Snowball stopword list to remove the stopwords in each narrative. Then, after counting the number of non-stopwords, we were able to calculate the proportion as described. Having the number of negative words and non-stop words in each narrative, we calculated the proportion using the function shown above.

We then grouped the data into “Student loan” and “Vehicle loan or lease” and plotted the proportions using a box plot.



We noticed that for both products, the outliers range from 8% to 50%, while most points are between 0% to 5%. The median proportion of negative words in “Student loan” complaints is 2.23%, and the median in “Vehicle loan or lease”: is 2.27%. The difference between the two proportions is only 0.04%. Visually, it is hard to tell if this tiny difference is statistically significant, so we conducted a hypothesis test to investigate if there is a difference in the means of the two groups.

Hypothesis Testing

Once we had calculated the proportions of negative words to the total amount minus stopwords, we proceeded to compare the data using Welch's t-test. This is because we found the variances in the two proportion datasets to be distinct. When this is the case, Welch's t-test can be used to compare the means of the two groups, even when the variances are unequal. Therefore, this test shows us if there is a statistically significant difference between the proportions of negative words in the complaints submitted for two groups of loans.

When performing the test, we found a statistically significant difference between the two groups, though a slight one. The p-value returned was 7.501e^-06, a number that is less than our predefined significance level of 0.05. This leads us to believe the observed difference is unlikely to be due to chance. While we reject the null hypothesis, suggesting the difference to be statistically significant, the means for

each group were very close in value. With this, we can conclude the difference to be minimally detectable from a human perspective and may have limited practical significance. This led us to explore whether we could create a classification model that could predict the loan type based on the proportion of negative words and other relevant variables.

Classification Models

Our first step in developing classification models is to use a few Naive Bayes models with different selections of predictors using a 70/30 train/test split. Our goal is to be able to classify narratives into one of the two types of loans without using any variables that are confounded with the product type. Furthermore, we would like to test if the proportion of negative words in a narrative is a helpful predictor of these categories. For the sake of clarity, we will use “Student loan” as the target or positive value.

The first model was designed more as a quality check than an actual attempt to create a model. The only input variable was “Issue,” which should uniquely define “Product.” In fact, this model does yield “perfect” results:

	uPredictions	
	Student loan	Vehicle loan or lease
Student loan	7012	0
Vehicle loan or lease	0	6928

The next model used variables from the original data but excluded variables that were obviously confounded, as well as variables with very high cardinality. This model shows moderate predictive power, although it has a very high rate of false positives.

	cPredictions	
	Student loan	Vehicle loan or lease
Student loan	6177	835
Vehicle loan or lease	3880	3048

The final model from this set used only variables calculated from the narratives. This includes the proportion of negative words, number of words, number of exclamation points, and others. It provides a prediction that is only nominally better than a model that predicts “Student loan” for all observations.

	tPredictions	
	Student loan	Vehicle loan or lease
Student loan	6435	577
Vehicle loan or lease	6277	651

The Accuracy, Precision, and Recall values for the three models are shown below:

Model	Accuracy	Precision	Recall
“Perfect”	1.00	1.00	1.00
Conservative	0.66	0.61	0.88
Text-based	0.51	0.51	0.92

Conclusion and Possible Next Steps

From Welch’s Two-Sample T-Test, we found a slight difference in the means of the two populations: Student loans vs. Vehicle loans or lease. The means were in terms of the proportion of negative words over non-stopwords in each narrative (percentage). Although theoretically, there was a difference in means, in a practical sense, it is too difficult for the human eye to detect. The models we did do, a conservative and text-based model, show that it is difficult to receive accurate classification results.

This finding gave us the idea that we could use different methods to differentiate the loans. For example, the use of a simple decision tree, which is easy to understand and interpret, could potentially help the process. People can follow along the tree’s path to see which nodes (classifying decisions) lead to a specific loan outcome. Moreover, we were thinking about adding predictors that could help differentiate loan types, like the use of capitalization (tend to suggest emotive language) and specific punctuation (exclamation points).

Works Cited

[1] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.
[2] Bing Liu, Minqing Hu, and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

Division of Labor

Xi: Quantifying the narratives(code and paper section), boxplot visualization

Alan: Basic statistics and visualizations (code and paper section), hypothesis testing (code and paper section)

Leah: Data cleaning (code and paper section), Naive Bayes classification model (code and paper section), timeline visualization, finding the data source

Jason: Motivation (paper section), decision tree model (code), conclusion and next steps (paper section)