FPM- T2- DAMR-Take Home Assessment

Shreya Singireddy

Emp ID: 31345

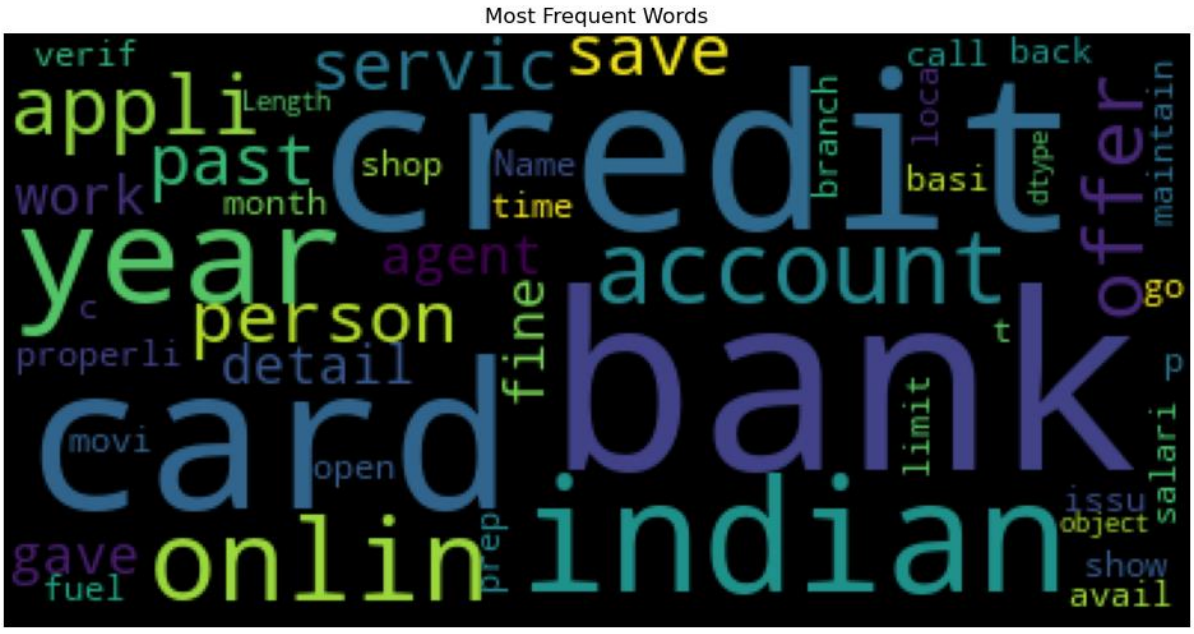
Exploratory Question: Attempting to explore the general sentiments and themes in the scraped Bank Reviews.

Our goal for this question is to explore the main discussion in these reviews and see their general opinions. We attempt to this using data scraping, data pre-processing, exploratory data analysis, sentiment analysis and explore the applicability of various supervised learning methods on the data scraped.

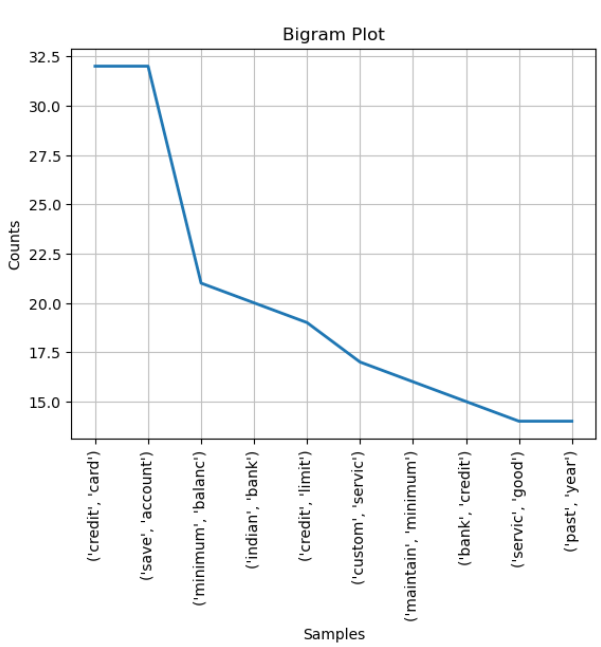
Data Summary:

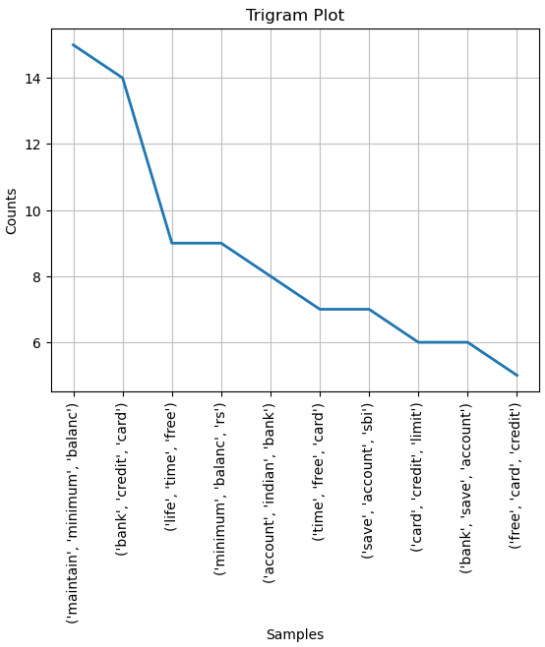
We scraped text review data from the website “BankBazaar.com” of various banks. This is done to potentially increase the diversity of the dataset. For our basic analysis and to diversify our understanding, we extract the bank title and cities the reviewers hailed from.

Before summarizing the data, we clean and pre-process our scraped text data using the ‘nltk’ package. We remove the stop words and remove special characters as well as numbers from the scraped text. Then we lemmatized the words to reduce the words to their base form and improve the accuracy of the text classification. Following is the patterns that we noticed from the generated plots:



The words ‘credit’, ‘bank’, ‘account’, ‘card’ are frequently used in the reviews given as per the word cloud. The pattern likely indicates the reviews are credit card customers or account holders.





In order to expand of the themes noticed in the word cloud, we generate bigram and trigram plots. The words ‘credit card’, ‘save account’, ‘minimum balance’ are some of the interesting themes from the bigram plot indicating the primary discussion among the reviews is regarding credit cards and saving accounts.

For refining it further we generate a trigram plot from which we get ‘maintain minimum balance’, ‘bank credit card’, ‘lifetime free’ as some words that are frequently used together. The trigrams further establish that it is likely the most common discussion among the reviews is related to the issues with minimum balance and credit cards provided by banks.

We also notice that from the data collected that most reviews are cities of Hyderabad, Bangalore, and Chennai.

Polarity and Subjectivity using Sentiment Analysis:

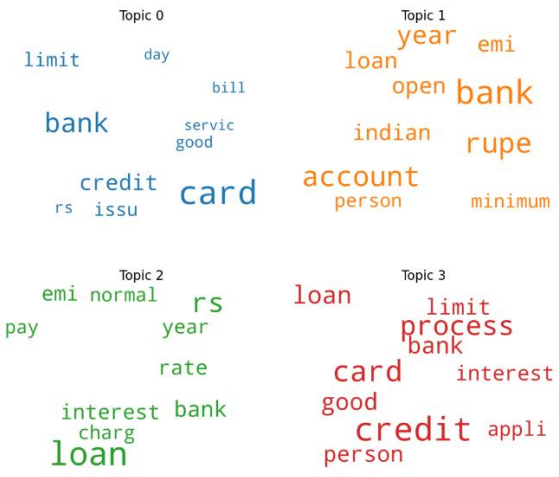
```
Sentiment(polarity=0.03762411347517731, subjectivity=0.4549290780141844)
```

The polarity and the subjectivity of the cleaned text were also tested. Polarity, which generally refers to the degree of negative or positive sentiment in the text, is typically measured on a scale from-1 to 1. As the polarity of this pre-processed data is 0.03762, it is indicated that the data is most likely neutral.

The subjectivity is for measuring the degree to which a piece of text expresses a personal opinion or an objective fact. It is typically measured on a scale from 0 to 1. As the subjectivity of this pre-processed data is 0.454929, it is indicated that the pre-processed data might be more of a neutral tone, which is surprising for a bunch of reviews.

Topic Modelling:

We have implemented Latent Dirichlet Allocation (LDA) to identify major topics in the collection of the documents. From the generated topics, we tried to check the dominant topics, the most representative sentence, and generated a word cloud for each topic. The dominant topic as per the generated topics and checking the most representative sentence for each topic is *"account" + "bank" + "good" + "servic" + "year" + "save" + "maintain" + "balanc" + "charg" + "salari"* indicating that the main theme might be regarding good service, maintaining minimum balance as well related charges for not maintaining the minimum balance and direct deposits for salaries. There might be generally a negative sentiment associated with it as the representative text that suggests that the customers' might have had bad experiences with it.

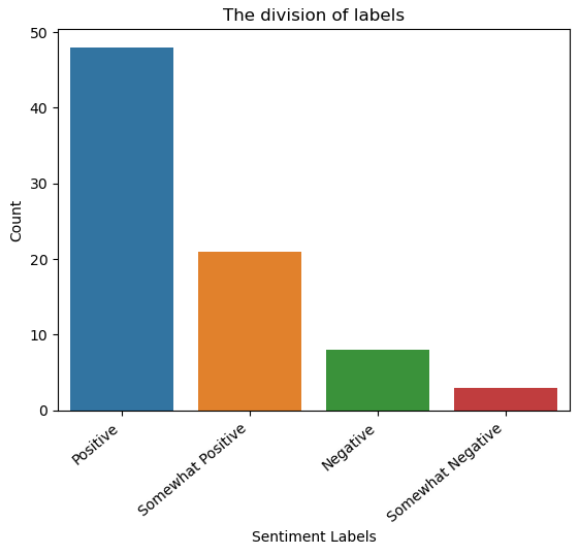


The Topic word cloud generally indicate that the majority of the discussion is regarding Bank Credit card, Loan interest rate, Loan EMI and application given.

Labelling Unlabelled data using Sentiment Analysis:

Before we perform any supervised learning, we need to label text data to train the machine learning model to predict the correct output for the new, unintroduced data. We will

attempt to label the data on the basis of sentiment analysis. We use Sentiment Intensity Analyzer in the 'nltk' library. It can identify the sentiment i.e., positive, negative, or neutral of a given text document by using lexicon-based approach to determine the polarity of the individual words and then aggregating the sentiment of the document. We binned the data into 4 categories: Positive, Somewhat Positive, Negative, Somewhat Negative. We notice that the data is imbalanced with way more positive reviews than negative reviews. The reason could be bias by the website to promote only positive reviews and pushing negative feedback to the end.



Dealing with Imbalanced Data while implementing Supervised Learning:

When the labels were applied to the data, the division of the data seemed to be imbalanced. This is a common problem with supervised learning and can lead to poor performance of the model especially with the minority classes. In our case, the values are overwhelmingly positive and very few negative values are present. We will attempt to balance the data while running various supervised learning techniques:

i. Using Logistic Regression:

For making a comparison with models where we attempt to balance the data, we will run a logistic regression model on the data as this model is the most sensitive to the imbalanced data and can lead to biased predictions. We get the test accuracy rate as **0.66**.

ii. Using Ensemble Models:

One of the primary approaches to deal with imbalanced data, is by applying ensemble model i.e., we combine multiple models to improve performance of the model and deal with the imbalance better. The test accuracy generated is **0.567**. As the accuracy of the model is lesser, this is an indicative that either the complexity of ensemble model is proving to be a disadvantage, or the model is overfitting.

iii. Using Random Under-sampling and applying Random Forest:

Under sampling can address the issues with imbalanced data by reducing the number of instances in the majority class to balance the class distribution. Since for the earlier ensemble model we might have faced the issue with overfitting, we will attempting random forest on the under sampled data. The Test Accuracy that was generated was **0.437.** The accuracy is lower than the earlier Ensemble Model as the it was likely trained on a smaller subset of data which led to the loss of information.

iv. Using Ensemble Model on the undersampled data

Ensemble model can leverage the strengths of multiple models to improve the overall models, hence, we try to run ensemble model on the earlier generated under sampled data. The Test Accuracy that was generated was **0.125**. The accuracy is far lesser than the earlier model as the data subset was smaller and the complexity of ensemble model prevented a more efficient training.

Best of the applied models and potential alternatives that could be applied:

Based on the models applied, we can say Ensemble model on the direct corpus might be a better option as it avoids the issues of undersampling and Logistic Regression's issues with unbalanced data.

Some of the other techniques that could be applied are overfitting, cost-sensitive learning. The best method to counter the imbalanced data is go back to the data collection stage and try to extract more data.