**Aritha Jayaratne**

ubmission date – 18/11/2023

**Text Classification Models for Customer Review Analysis.**

Contents

[Introduction. 2](#_Toc151199188)

[Importance of analyzing customer reviews. 3](#_Toc151199189)

[The objective of implementing a text classification model. 3](#_Toc151199190)

[Challenges. 4](#_Toc151199191)

[Opportunities. 4](#_Toc151199192)

[Data Collection. 5](#_Toc151199193)

[Methodology. 6](#_Toc151199194)

[Insights and Interpretation. 6](#_Toc151199195)

[Sentiment Analysis. 11](#_Toc151199196)

[Latent Dirichlet Allocation (LDA) Model. 12](#_Toc151199197)

[Model selection – Naïve Bayes text classification model. 13](#_Toc151199198)

[Model training and Evaluation. 13](#_Toc151199199)

[Results Evaluation. 14](#_Toc151199200)

[Recommendations. 15](#_Toc151199201)

[**Product and Service Enhancements.** 15](#_Toc151199202)

[Marketing and Customer Engagement Strategies. 15](#_Toc151199203)

[**Implementation Plan.** 16](#_Toc151199204)

[Conclusion. 17](#_Toc151199205)

[References. 18](#_Toc151199206)

# Introduction.

In the modernized digital world, the banking industry has been revolutionized with various new technologies. Customer feedback and customer reviews stand as one of the most valuable assets for any financial organization. Understanding customer emotions and expressions through their reviews provides banks with valuable insights about their customer services and helps them improve their loyalty and customer satisfaction. The growth of online reviews has paved the way for companies using analytical techniques and machine learning algorithms to extract meaningful patterns from the customer reviews from a huge amount of data available. The report presents a comprehensive analysis of customer reviews from various banks in the United States of America and the different models and data analysis techniques used in the process.

The main objective of the report is to gain valuable insights from the customer reviews available and use natural language processing and machine learning algorithms to identify the sentiment of the data set and predict the sentiment of future reviews using the proposed model. By doing these techniques the team aim to provide insights and recommendations to the banking institutions who are seeking to implement the model into their organization process and improve their customer experience in the future.

This report consists of the data gathering steps, cleaning and model development steps used to achieve these insights. This explains in depth the importance of the naïve bayes model suitability for banks or any organization to perform sentiment classification using text. The analysis will highlight the main observations and the patterns in the customer review text and this will help to analyze customer reviews separately for each bank by doing data processing. By the power of machine learning the team aims to provide informed strategic decisions and how to improve the customer satisfaction in United states banks.

## Importance of analyzing customer reviews.

* **Insight into Customer Preferences and Needs:** Customer reviews provide a direct window into the preferences, needs, and expectations of the target audience. By analyzing reviews, a marketing company can gain valuable insights into what customers appreciate, dislike, and desire, helping tailor marketing strategies to better align with customer expectations.
* **Identifying Strengths and Weaknesses**: Reviews often highlight both the strengths and weaknesses of products or services. By carefully analyzing this feedback, a marketing company can identify key selling points to emphasize in marketing campaigns and address any weaknesses in subsequent product/service iterations.
* **Building Credibility and Trust:** Positive customer reviews serve as powerful testimonials that can be incorporated into marketing materials. Sharing real customer experiences helps build credibility and trust with potential customers, making marketing efforts more persuasive and authentic.
* **Refining Target Audience Segmentation:** Customer reviews can provide insights into the demographics, interests, and behaviors of the target audience. This information is crucial for refining and optimizing target audience segmentation in marketing campaigns, ensuring that promotional efforts reach the right people with the right message.

## The objective of implementing a text classification model.

* **Efficiency and Scalability:** One of the primary objectives is to improve the efficiency of handling large volumes of unstructured text data. Manual analysis and categorization of text can be time-consuming and prone to errors. Text classification with machine learning enables scalability, allowing companies to process vast amounts of text data in a timely and cost-effective manner.
* **Automation of Business Processes:** Implementing a text classification model aims to automate the process of organizing and understanding text data. By automatically assigning predefined categories or labels to text, businesses can streamline their workflows and reduce the reliance on manual labor for tasks such as sorting through emails, analyzing survey responses, or categorizing social media comments.
* **Data-Driven Decision Making:** Text classification helps organizations make data-driven decisions by extracting valuable insights from unstructured text data. By categorizing text into relevant topics or sentiments, companies can identify patterns, trends, and critical information that can inform strategic decision-making.
* **Real-Time Analysis:** The ability of text classification models to analyze text data in real-time is crucial for addressing time-sensitive situations. For example, in PR crises on social media, real-time analysis allows companies to quickly identify brand mentions, sentiments, or emerging issues and take immediate action to mitigate potential risks.

## Challenges.

* **Fake Reviews and Trust Issues:** The rise of online reviews has also led to the proliferation of fake reviews. Businesses may engage in unethical practices to boost their reputation, making it challenging for consumers to trust online reviews.
* **Bias in Reviews:** Some reviews may be biased or influenced by personal preferences, experiences, or even competitors. Sorting through subjective opinions to extract valuable insights can be difficult.
* **Managing Negative Reviews:** Negative reviews can significantly impact a business's reputation. Small businesses may find it challenging to manage and respond effectively to negative feedback, especially when dealing with limited resources.

## Opportunities.

* **Global Reach for Small Businesses:** E-commerce allows small businesses to reach a global audience without the need for a physical presence in various locations. This provides an opportunity for growth and expansion.
* **Building Authentic Reputation:** Positive reviews from genuine customers can help small businesses build an authentic reputation. This authenticity is crucial in a market where consumers value transparency and honesty.
* **Customer Engagement and Loyalty:** The example of the CEO reaching out for feedback highlights the opportunity for small businesses to engage directly with customers. This engagement can lead to increased customer loyalty and repeat business.

# Data Collection.

The dataset used in the report and the model was procured from Kaggle. The data set has customer reviews from all the banks which are operating in United States of America from the years 2017 to 2023. The data set had nearly 20,000 data points and this was very much useful in the exploratory data analysis, and the model building steps. According to the authors of the data set the customer reviews were collected by doing web scraping and doing live online surveys and compiling them all together (Viswanathan, 2016).

The data cleaning steps used in cleaning the data set was removing null values and rows. Cleaning the text for unusual characteristics even though this was again done in the preprocessing stages of the model building. Also, the data column was broken down into day, month, and year to make it easy for the exploratory data analysis to derive valuable insights from the data set. The EDA will be explained further in the next part of the report for a comprehensive understanding.

# Methodology.

**Text processing: Exploratory data analysis**

To analyze and summarize the main characteristics for our dataset we used exploratory data analysis where we fully understand, identify patterns, relationships, and anomalies. We came up with basic information such as number of reviews by banks in USA, distribution of star ratings for banks, average star rating for months and years, top review location for banks, flow of reviews within 7 years and top 5 banks with highest star rating. (Chaerul Haviana et al., 2023)

**Feature extraction:**

Term frequency to measure how often a term appears in our document.

**Model selection: Naïve bayes**.

Naïve bayes is a popular machine learning algorithm particularly for text classification, spam filtering, sentiment analysis and topic categorization. It is also good for dealing with high dimensional data’s and can be easily implemented.

# Insights and Interpretation.

A graph of reviews by bank

Description automatically generated

Figure 1.1: Reviews by Banks

The above figure 1.1 shows all the banks with reviews stating from 0 to more than 1500. There are six banks which have managed to get over than 1000 reviews, but bmo-harris-bank has got more than 1500 reviews having the highest reviews compared to other banks. Banks such as Bofa gift card and usaa banking also have around 1400 reviews. The lowest number of reviews are obtained by banks such as bbt bank, armed forces bank and furthermore.

A graph with a bar

Description automatically generated

Figure 1.2 Distribution star rating

From the figure 1.2 we obtain that the reviews we receive from customers are mostly negative as they have highly rated the banks by giving 1 star. The marketing team can focus on the negative issues of customers as the rating clearly shows that the banks performance is not satisfactory for customers.

A graph of numbers and lines

Description automatically generated

Figure 1.3: star rating by month and year

The figure 1.3 reveals the month and year where banks get reviews. In the year 2018 the banks receive high reviews in the month of May, moving to the year 2020 the reviews get high in the month of July and August. In the year 2021 the review again shifts to month of May and June. This reveals the trend of receiving reviews by banks each year differs to the months.

A graph of blue bars

Description automatically generated

Figure 1.4: Top location for Reviews

The above figure 1.4 Shows the top location for the reviews of banks in United Stated of America. The highest review received if from Chicago followed up by Los Angeles and Houston. This information also reveals a lot of customers are from this location who use the banking service.

A graph with red line

Description automatically generated

Figure 1.5 Number of reviews over the years

The figure 1.5 revels the trend for reviews for 7 years. The reviews had consistent growth till 2019 reaching the top. From the year 2019 we see a significant decrease in reviews till 2023. The reason for highest reviews received in 2019 could be the Covid-19 pandemic. We also have a lot of negative reviews which could indicate the improper measurement for giving services during covid. With the proper implication of covid 19 and lockdown by governments we could see the result of negative reviews getting lesser and lesser.

A screenshot of a graph

Description automatically generated

Figure 1.6: Top 5 Banks with 5-star rating

The figure 1.6 reveals all the banks with all the positive rating of 5 stars with Merrick bank leading in the top spot, followed by Wells Fargo, bofa, chase bank and navy federal. The chart gives important information as the banks have good, satisfied customer base and marketing team can use information from these banks to come up with new plans to attract customers.

## Sentiment Analysis.

After a thoroughly doing an exploratory data analysis on the data set the team performed a sentiment analysis on the customer review column of the data set and found out the following insights relating to the banks in USA. Based on the sentiment analysis done using emotions, two models were developed for the marketing team to predict future insights and to find needed solutions for the problems that were identified. Based on the results predicted by the sentiment analysis a graph was plotted.

A graph of different colored bars

Description automatically generated

The graph shows the distribution of the customer reviews as sentiments. The x axis representing the sentiment score and the y axis is representing the percentage of reviews on the data set column customer reviews. The Sentiments consist of 10 terms as shown on the graph. The negative term has the highest score on the graph and the second highest is the positive column. With both trust and fear being on the same level on the review count. So based on the graph the team concluded that the sentiment of the text reviews is mostly negative customer reviews on the data set.

This is not necessarily bad. Sentiment analysis is not a perfect method I differentiating the text reviews. It is possible for the algorithm to incorrectly categorize the sentiment review. And, in general some banks might be more likely to gain negative results or reviews than positive emotions.

## Latent Dirichlet Allocation (LDA) Model.

According to the cleaned bank data set the team used a Latent Dirichlet Allocation (LDA) model with 10 topics to get insights about the Bank customer review column and used these insights to provide recommendations to the marketing company. The 10 topics are aligned with the key areas aligned with the interest of bank customers. The highest frequency terms suggest by the model provides valuable analysis on the customer concerns and experiences the customer experienced throughout the targeted years. According to the terms present in the graph, the main areas the Marketing teams of banks should investigate.

* Accounts and Services: Bank account, Customer services, Credit cards, Mortgage, Loans.
* Fees and Chargers: Credit, ATM, Pay, Payment, Service charge, due.
* Online banking: Online banking, Bill pays, Mobile, Banking, Website.
* Customer support: Service, Customer, Called, Help, Phone, Time.
* Fraud and security: Fraud, Security, Never, New.

A screenshot of a graph

Description automatically generated

The marketing team should focus on improving these areas according to the model and based on the words suggested. Also, another useful insight to mention is the number of 1-star ratings in the data set. This shows that most of the customers and their reviews on the banks are negative reviews. The team should focus more on how to improve these areas on the long run.

# Model selection – Naïve Bayes text classification model.

Reason for selecting the naïve bayes model text classification is because the model is known for its computational efficiency and the quick training and prediction times safes up a lot of time for the marketing team. And the model is a straightforward algorithm, and I is easy to implement into any system or an organization. Also, the model has features such as NLP the ability to handle categorical data. The model is designed to handle and compute even limited number of data and the features help the model to move out irrelevant features.

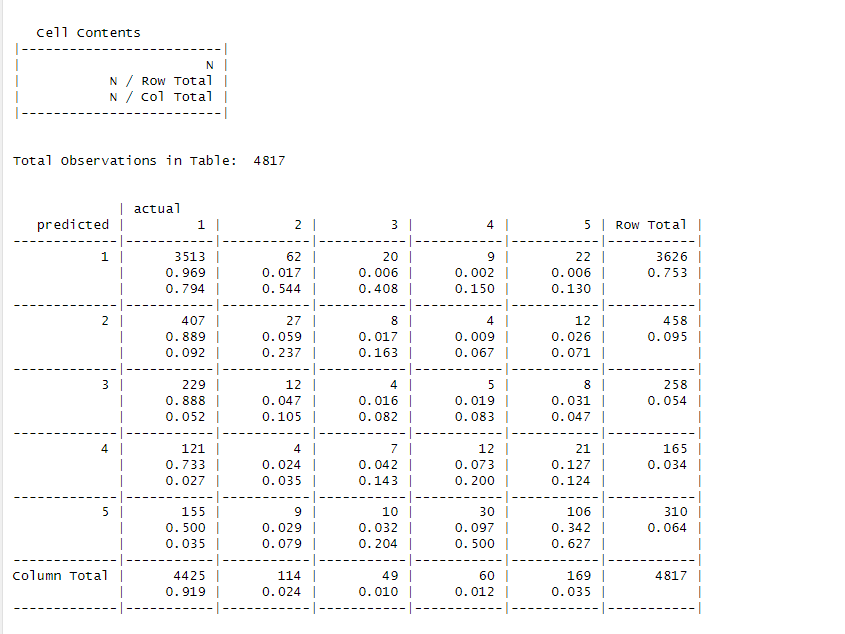
# Model training and Evaluation.

The model training process can be summarized into several key steps. Create a customer review corpus and preprocess the text by turning the text into lower cases, remove numbers, punctuations and the whitespaces. The cleaned corpus is transformed into a Document term matrix (DTM) by using the tm package which is available in R language which represents the frequency of each word in the document (Sang-Hwa Lee, 2023).

The dataset is split into as training and testing data set. 75% data into training and 25% into testing data respectively and also creating a DTM for each data. The dependent variable which is used in this scenario is the star rating column that is the data set. The code identifies the most frequent terms present in the training set using the FindFreqTerms function. The resulting frequent term is used to filter the training data and the testing data.

Then in the feature selection process, a logic is created to convert the term frequency counts in the binary form, Yes or No is used to indicate the presence and the absence of each term in the document. The model is then trained using the Naïve bayes function which is available in R language. The trained model is used to predict the start ratings on the testing set and the results are plotted using a CrossTable function.

## Results Evaluation.



The above table shows the predictions for the testing data set which had 4817 observations. The Naïve bayes model is performing well on the testing data set which is used here. The overall prediction of the model is 75% which is more accurate than the chance accuracy which is 33% (Zhang et al., 2019). The model performed well on the individual classes. For an example if you look at the 1 star ratings it has 97% accuracy. The model is a good fit for the data set and can predict accurate results for the class of the reviews.

The predictions generated by this model can help to identify positive ore negative reviews based on the predicted star ratings. These predictions can be used be banks to act and improve certain sectors of the banks.

# Recommendations.

Considering the bits of knowledge acquired from the Naive Bayes model, the bank can make a few significant suggestions to further develop consumer loyalty. These suggestions can be comprehensively classified into two groups: product and service enhancements, and marketing and customer engagement strategies.

## **Product and Service Enhancements.**

* Address common customer complaints: The model can be used to determine which issues come up most frequently in evaluations from customers. After that, the bank can decide which of these complaints to prioritize, including cutting down on wait times for customer assistance or fixing common technical problems.
* Develop new products or services: The model can be used to determine client preferences and needs that aren't being satisfied by the bank's current offers. The bank can then utilize this data to create new offerings that are more likely to be liked by its clientele.
* Personalize product recommendations: By examining client evaluations, the model may spot trends in their preferences. Customers can then receive personalized product recommendations based on this information. For instance, a client who regularly communicates.

## Marketing and Customer Engagement Strategies.

* Target marketing campaigns: The model can be applied to identify consumer groups with similar requirements or preferences and to segment customer reviews based on sentiment. The bank can then direct marketing campaigns towards these demographics to make sure they are exposed to pertinent and interesting content.
* Track sentiment on social media: The model can be used to track user sentiment on social media sites. This can assist the bank in spotting new problems or patterns and taking preventative action.
* Gather customer feedback: The approach can be used to evaluate reviews left by clients and pinpoint areas in which the bank's client communications need to be improved. The bank might utilize this data, for instance, to produce more FAQs or more thorough product descriptions.

# **Implementation Plan.**

The Naive Bayes model can be integrated into the bank's existing systems in a variety of ways.

* Data collection and storage: The bank must set up a procedure for gathering and keeping client testimonials. There are other ways to store this data, including databases and data lakes.
* Data pre-processing: Before the Naive Bayes model can use the customer reviews, they must first be pre-processed. This involves stemming or lemmatizing words, reducing text to lowercase, and eliminating punctuation and special characters.
* Model training and evaluation: A representative sample of customer reviews must be used to train the Naive Bayes model. To make sure the model is generalizing well, it should then be assessed on an independent test set.
* Model deployment: Model maintenance and monitoring: The Naive Bayes model needs to be regularly retrained and checked for changes in user feedback.

Example:

Ally Bank has used a Naive Bayes text categorization technique has been employed by digital bank Ally Bank to enhance customer support. The model is used to categorize customer evaluations automatically and spot possible problems. This enables the bank to proactively resolve client issues and stop them from escalating (Domingos, 2015).

# Conclusion.

In conclusion, with implementation of natural language processing, machine learning algorithms and advanced statistical techniques the banking industry has been transformed into an industry which uses advanced technology to analyze customer data. The comprehensive analysis done on the customer reviews of the Banks in United states has given valuable insights into customer experiences, sentiments and the strengths and weaknesses which banks should address in the short and long run. Using technology such as naïve bayes model for analyzing customer reviews has proved to be an efficient and effective way to process a large number of unstructured text data.

The exploratory data analysis revealed some key points to investigate the bank data set such as the distribution of the star ratings, analyzing the trend over the years, and the top locations for reviews. The sentiment analysis focus on the negative reviews and the LDA model focus on the areas where the banks should focus more based on the model. These areas are critical for a banking institution. The Naïve bayes model predicts positive results based on the testing data set and this could lead to making strategic decisions within the organizations.

The report shows some of the major roles of AI and machine learning techniques that will be used to shape the future of the banking industry as well as any organization which uses AI and ML for decision making and to continuously improve customer experiences.

# References.

* Chaerul Haviana, S.F., Mulyono, S. and Badie’Ah (2023) ‘The Effects of Stopwords, Stemming, and Lemmatization on Pre-trained Language Models for Text Classification: A Technical Study’, *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Electrical Engineering, Computer Science and Informatics (EECSI), 2023 10th International Conference on*, pp. 521–527. doi:10.1109/EECSI59885.2023.10295797.
* Chingmuankim and Jindal, R. (2022) ‘Classification and Analysis of Textual data using Naive Bayes with TF-IDF’, *2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Electrical, Control and Instrumentation Engineering (ICECIE), 2022 4th International Conference on*, pp. 1–9. doi:10.1109/ICECIE55199.2022.10000309.
* Zhang, L., Jiang, L. and Li, C. (2019) ‘A discriminative model selection approach and its application to text classification’, *Neural Computing and Applications*, 31(4), pp. 1173–1187. doi:10.1007/s00521-017-3151-0.
* Sang-Hwa Lee (2023) ‘Text Classification of Mixed Model Based on Deep Learning’, *Tehnički Glasnik*, 17(3), pp. 367–374. Available at: https://research.ebsco.com/linkprocessor/plink?id=35465ef9-cbbd-3eaa-9cae-d18d4e825629 (Accessed: 17 November 2023).
* Syeda, F.S. (2020) *Sentiment Analysis of Financial News with Supervised Learning*. Linköpings universitet, Institutionen för datavetenskap. Available at: https://research.ebsco.com/linkprocessor/plink?id=cde2250b-219e-3454-85a2-c7d89ad5dfbc (Accessed: 17 November 2023).
* Bali, R. (2016) *R*. Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=55a85a65-1b90-34c9-9b60-227ec14721a4 (Accessed: 17 November 2023).
* Viswanathan, V. (2016) *R : recipes for analysis, visualization and machine learning : get savvy with R language and actualize projects aimed at analysis, visualization and machine learning / Viswa Viswanathan [and three others]*. Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=159c7132-49d6-3672-a9d6-09ea36e1d077 (Accessed: 17 November 2023).
* Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.