SAN DIEGO STATE UNIVERSITY

BDA 594 Big Data Science and Analytics Platforms

Financial Product Complaint Research

[Project Website]

Nischita Biradar

Problem Statement

Effective resolution of consumer complaints regarding products and services of financial institutions, is paramount for enforcing financial regulatory standards and upholding best practices. Such cases regarding financial products require more urgency to prevent adverse impacts from negligence of acting financial institutions.

The CFPB (consumer financial protection bureau) is a US government agency that aims to process such consumer complaints and ensures that financial institutions are in compliance with financial law. The core functions of the agency include providing monetary relief to affected customers, informing them on financial standards and concepts, and overseeing transactions and financial activity between financial institutions and consumers. Consumer complaints are stored along with other information regarding cases in a publicly available database, that contains the textual data represented through fields that indicate the consumer's original written complaint as well as the corresponding company's response to the complaint.

Our aim is to conduct exploratory data analysis on the consumer complaint dataset extracted from the CFPB website, last updated during September 2023. Insights we wish to find include examining the most prominent categories of consumer complaints by various stratifications. Textual analysis such as finding the most prominent composition of topics and words based on a latent dirichlet allocation (LDA) model and word cloud will be carried out. We will also formulate a spatial temporal display of the concentration of complaints by region and time period. Our analysis will establish a comprehensive narrative surrounding the nature and key text trends shown from complaints, which provides real world inferences on risks, operational issues and consumer attitudes associated with financial products.

The CFPB processes on average 25,000 complaints per day, and has received over 4 million complaints historically regarding various issues and products. Examined in the dataset, we see that there are more than 10 major topic types for the complaint cases within the "Product" column. The automated classification/segmentation of complaints into distinguishable clusters of topics may contribute to faster response times and better resource allocation. A classification model that classifies complaints based on their topic type shall be constructed for high accuracy multi-labeling. In our predictive modeling portion, we will build a Long Short-Term Memory (LSTM) model, which is a recurrent neural network model through the TensorFlow package on Python.

Literature Review

There are several studies that have utilized algorithms such as XGBoosting, Support Vector Machines, LSTM, Bidirectional LSTM (BLSTM), Convolutional Neural Networks (CNN) and Naive Bayes in text classification, specifically consumer complaint data.

A study (D.O. Oyewola et al, 2023) on consumer complaint data extracted from the CFPB, utilized a Two-Stage Residual One-Dimensional Convolutional Neural Network (TSR1DCNN) to classify consumer complaint data into different product categories. We see another study (Pramod Kumar Naik et al, 2022) that takes a similar multiclass classification perspective with the same consumer complaint data, where it built various machine learning models such as Naive Bayes, Decision Trees and SVM. It transformed the complaint data into vectors of unigrams and bigrams combinations- to be used as the inputs respectively.

Studies regarding text classification are shown to commonly use BLSTM or other related hybrid models involving a BLSTM layer. This is evident in a recent study (U. B. Mahadevaswamy et al, 2023) where a BLSTM model was utilized for classifying the sentiments of over Amazon product reviews. It utilized word embedding, BLSTM and Dense Layer as the core components of the neural network model for binary classification of the text into positive and negative sentiment.

There are a multitude of studies that have conducted text mining, namely LDA topic modeling. However, there has only been one study that applied LDA modeling onto the CFPB complaint data. In the study (Kaveh Bastani et al, 2019), an LDA model was built to accept a term document matrix of unique words within the corpus and its frequency in each document. The time trend of the the top 12 most prominent topics that were formed from the LDA model were also displayed to show the temporal distribution of the most prominent topics on a monthly basis.

Database management and Data Process Procedure

The dataset for our project is sourced from the Consumer Complaint Database from the Consumer Financial Protection Bureau (CFPB). Complaints are submitted by consumers after following a defined process, which includes details such as product type, complaint, consumer narrative, and the company's response. Among these complaints, only the ones which are sent to companies for a response are eligible to be a part of the database.

The fields in the dataset consist of "Date received", "Product", "Sub-product", "Issue", "Sub-Issue", "Consumer Complaint Narrative", "Company public response", "Company", "State", "ZIP Code", "Tags", "Consumer consent provided?", "Submitted via", "Date sent to company", "Company response to consumer", "Timely response?", "Consumer disputed?", "Complaint ID".

The dataset is updated regularly, For purposes of consistency, we use the dataset downloaded on 13 September 2023 in all steps of our project and visualizations. This database can be accessed in various ways, CSV or JSON formats, Open Data API, etc. We utilized the CSV format for our analysis. The CSV file is accessed using Microsoft Excel.

The data preprocessing steps we undertook for the consumer complaint database data involved several key actions to clean and prepare the data for analysis:

Filtering out NA values: Removing any missing or null values (denoted as 'NA') from the dataset.

Product and issue categorization: Consolidating various product types and issues into broader categories for clarity. For Example: Replaced ‘Credit reporting, credit repair services, or other personal consumer reports’ with ‘Credit reporting’

Removing Special Characters: Eliminated non - alphanumeric characters (like @, #, $, etc.) which are irrelevant for our analysis.

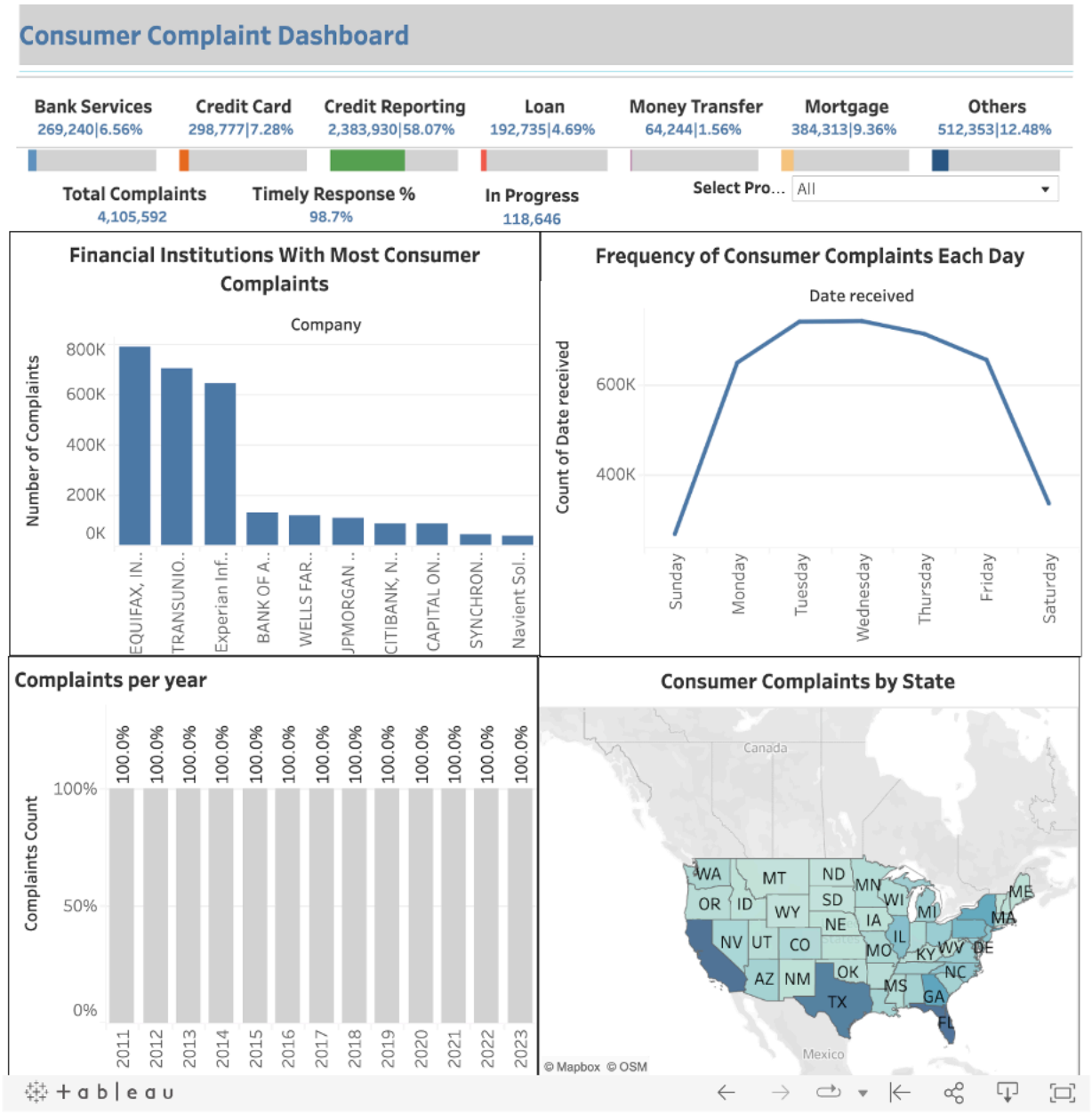
Removing Stop words: Stop words are common words (like 'the', 'is', 'at', 'which') that do not add significant meaning to the text were removed for creating word clouds and text analysis.

Lemmatization: Reduced words to their base or root form. For example, 'running', 'ran', and 'runs' would all be lemmatized to 'run'.

Tokenization: Broke down text into smaller units called tokens, typically words or phrases. Tokenization is a fundamental step in text preprocessing, as it converts text into a format that can be easily analyzed and used in various natural language processing (NLP) applications.

Each of these steps played a crucial role in preparing the dataset for in-depth analysis, ensuring that the data is clean, uniform, and structured for effective processing and insight generation.

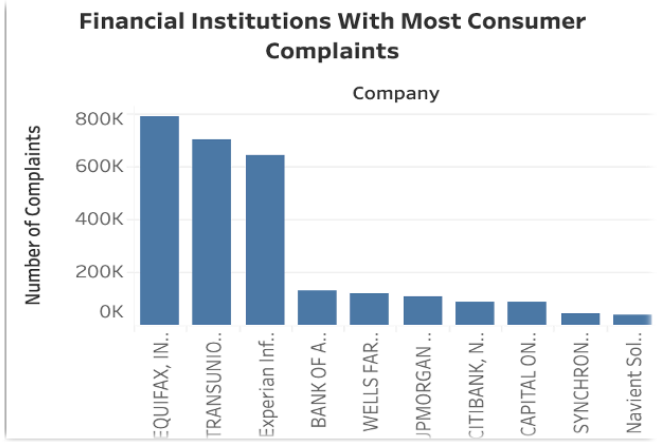
Data Analysis and Visualization Results



This consumer complaint dashboard provides valuable insights into the correlation between consumer complaints and other attributes. At the top, we have Key Performance Indicators (KPIs) for each product

category and total complaints. To enhance interactivity, there's a dropdown menu allowing users to filter visuals by product category.

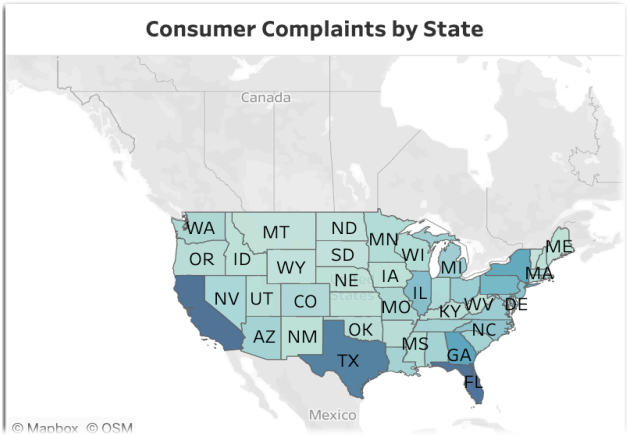
Financial Institutions With Most Consumer Complaints



Dominant Player with Complaints: EQUIFAX, 'TRANSUNION INTERMEDIATE HOLDINGS, INC.' and 'Experian Information Solutions Inc. have significantly higher complaints than the Others. These are likely credit reporting agencies The chart shows a clear descending order of complaints, with a steep drop after the first three companies, which could indicate a higher concentration of complaints within a specific segment of the financial industry.

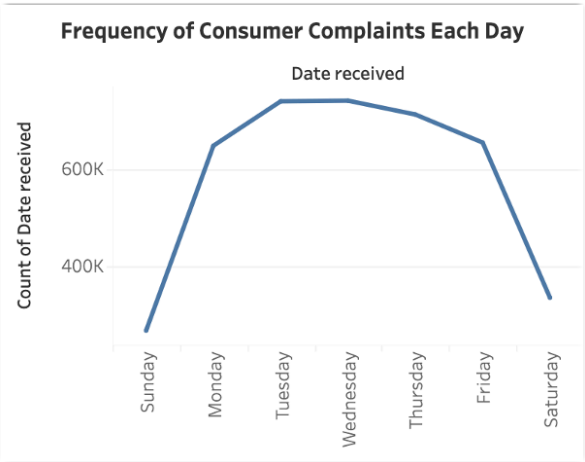
In summary, The stark differences in complaint volumes can help regulatory bodies, consumer protection agencies, or the companies themselves to prioritize areas for improvement in customer service, compliance, and product offerings.

Consumer Complaints by State



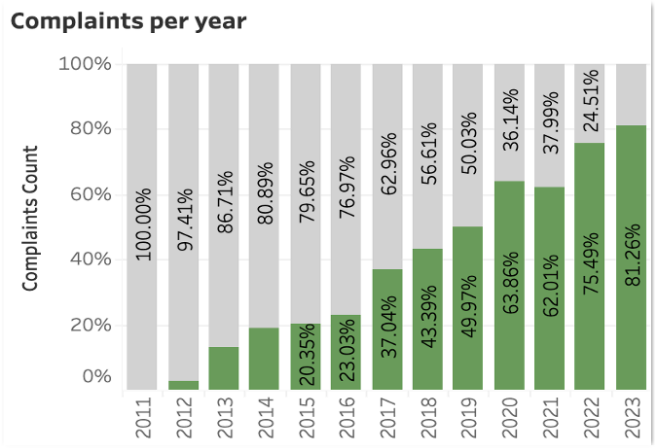
This visual is a choropleth map titled "Consumer Complaints by State." This type of map uses differences in shading or coloring within predefined areas, in this case, states, to indicate the volume of consumer complaints. The darker shades on the map suggest a higher volume of complaints. For instance, California (CA) and Texas (TX) have the darkest shades, indicating they have the highest number of consumer complaints among all states. States with Fewer Complaints: States such as Wyoming (WY) and the Dakotas (ND and SD) are much lighter, implying fewer complaints. This could be due to smaller populations or possibly fewer financial institutions or transactions. There seems to be a correlation between the population of a state and the number of complaints. More populous states like New York (NY) and Florida (FL) also have darker shades, supporting this observation.

Frequency of Consumer Complaints Each Day



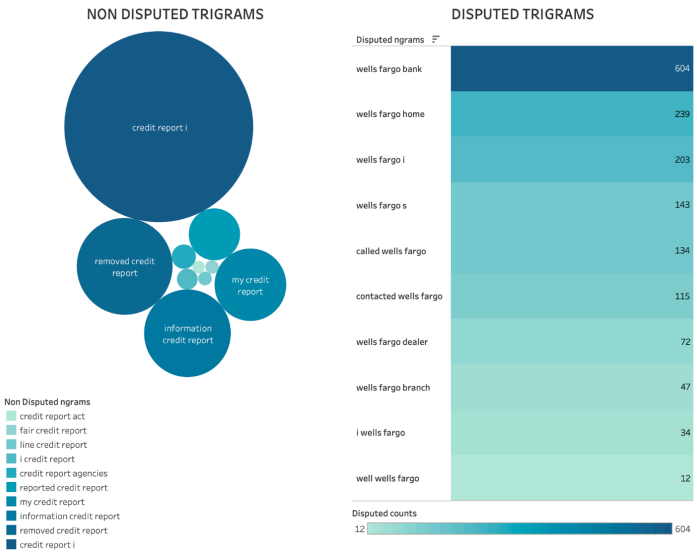
This is a line graph titled "Frequency of Consumer Complaints Each Day." It represents the count of consumer complaints received on each day of the week. There is a clear peak in the middle of the week, with Wednesday being the day with the highest number of complaints. This suggests that consumers are more likely to file complaints on this day. The number of complaints drops significantly on weekends, with Saturday showing the fewest complaints. This could be due to reduced hours of operation for financial institutions and customer service centers, or simply consumer behavior being different on weekends. Monday also shows a high number of complaints, which might indicate that issues that occurred over the weekend or were not addressed the previous week are reported as the new week begins. This information could be invaluable for customer service operations, indicating the need for more staff or resources midweek, especially on Wednesday.

Complaints Percentage Every Year



This is a stacked bar chart that depicts the percentage of complaints per year for different product categories. A clear trend of decreasing complaint rates is observed for every product, except for credit report agencies, reflecting consumers' sensitivity to credit information accuracy and the improvement in financial institutions and products over years for other products. Collectively, these insights help us understand complaint dynamics, guide resource allocation, and prioritize efforts towards specific financial sectors to enhance consumer satisfaction.

Trigram Analysis

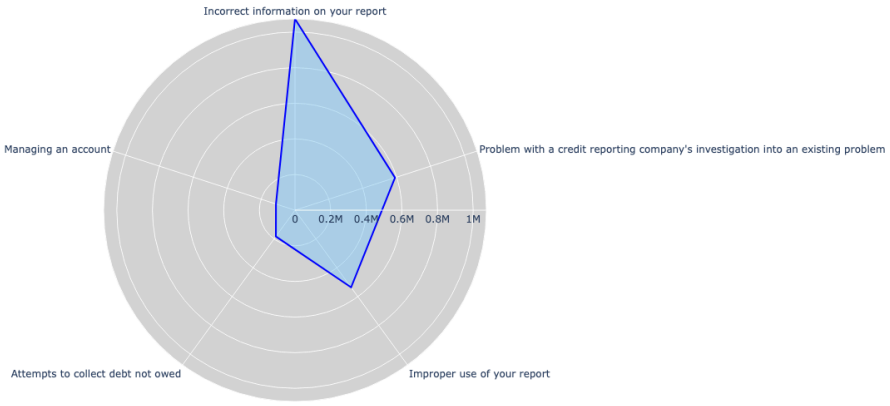


In our analytical exploration of consumer complaint narratives using a Tableau dashboard, we employed a distinctive approach to discern and visualize prevalent themes. Our analysis focused on extracting the most common trigrams - groups of three consecutive words - from the narratives, with a bifurcation based on the nature of the complaints: disputed and non- disputed. It is important to understand that all the complaints in our universe of data already have a response from the company, since that is a prerequisite to be stored in the database. However, some of these responses are then disputed, which creates a recurring, back-and-forth issue, which gives it the title ‘disputed’ in the consumer disputes field.

The left side has the top 10 non disputed trigrams, these trigrams predominantly revolved around credit reporting issues, with recurring combinations like ('credit', 'report', 'I'), ('information', 'credit', 'report'), and ('removed', 'credit', 'report').Since complaints relevant to credit reporting come under non-disputed it implies that while these issues are common among consumers, most of these are not severe enough to be disputed. So, maybe the complaints were resolved and customers were satisfied. Indicating that credit reporting agencies are working efficiently at resolving these complaints. Right side shows the top 10

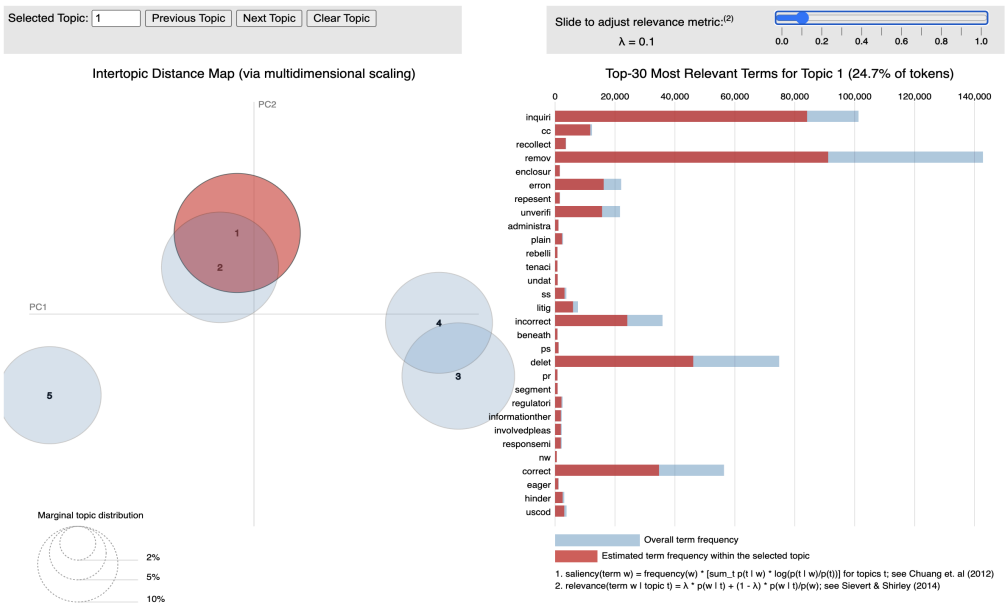
disputed complaints. Intriguingly, these trigrams were heavily centered around interactions with Wells Fargo, as seen in combinations like ('wells', 'fargo', 'home'), ('wells', 'fargo', 'bank'), and ('contacted', 'wells', 'fargo'). Wells Fargo has the fifth highest number of complaints but the trigrams are addressed towards Wells Fargo. This could be due to poor customer service or product service. We also see that apart from Wells Fargo there are a variety of words like bank, dealer, branch indicating that these complaints are not related to only a single area of operation. It is spread across multiple areas of the bank's functions.

Radar Chart



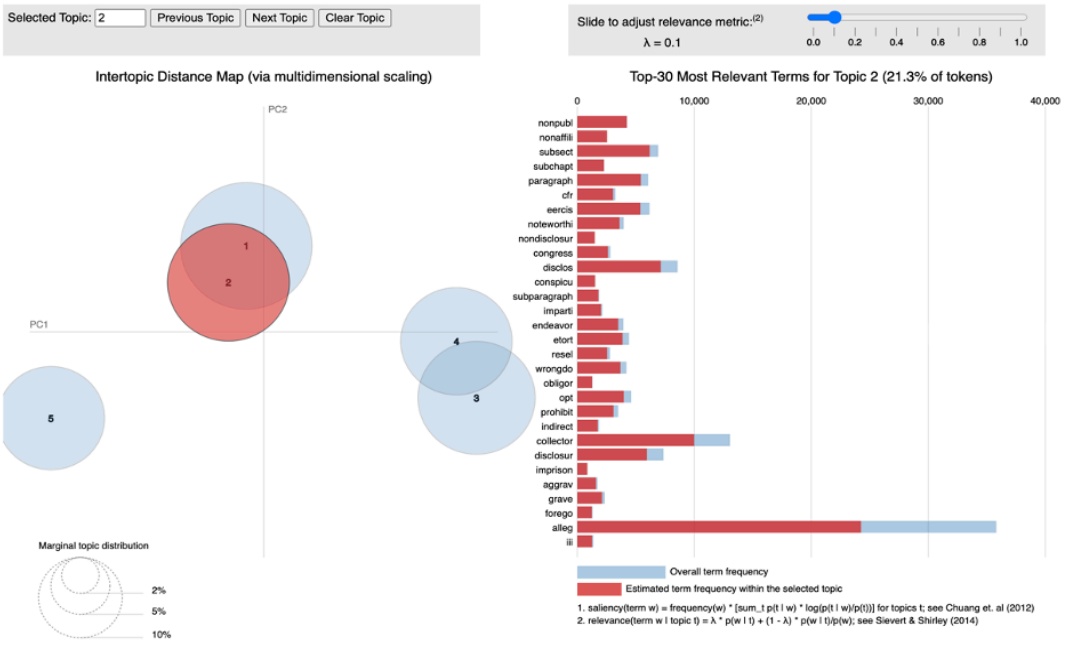
This is a radar chart focusing on the most common issues in our dataset. It gives us a comparative view of the top 5 issues. Each axis of the chart represents one of these key issues, with the length of each axis corresponding to the issue's prevalence.The area of the shape gives us the overall magnitude of all of these issues. Each axis of the chart represents a different issue and the radius of the chart displays the instances in millions. The higher the area of the shaded region towards a particular issue, the greater the occurrence of that issue. In our case, incorrect information on your report is the most common issue with about one million instances whereas managing an account has only about a hundred thousand instances.

Latent Dirichlet Allocation

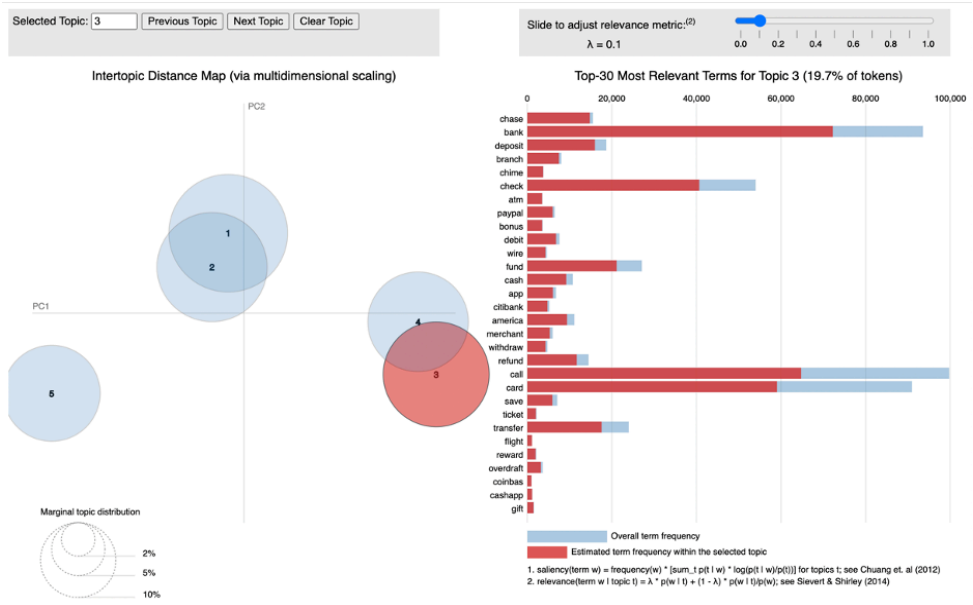


We choose to build an LDA topic model composed of 5 topics, to further aggregate the 9 main product groups found within our dataset of complaint cases to see if we can generate a distinct cluster of 5 new topics. In the output as shown, we see the topic clusters plotted on a 2-dimensional axis through their principal components, and the distribution of words within each. The relevance metric in the sliding scale determines how the terms are ranked. A high value will rank the terms that are the frequent within a topic first while lower rankings will give more weightage to unique terms in a topic. We shall set our value to 0.1 to see the unique terms for each topic, which could reveal their characteristics and prominent trends.

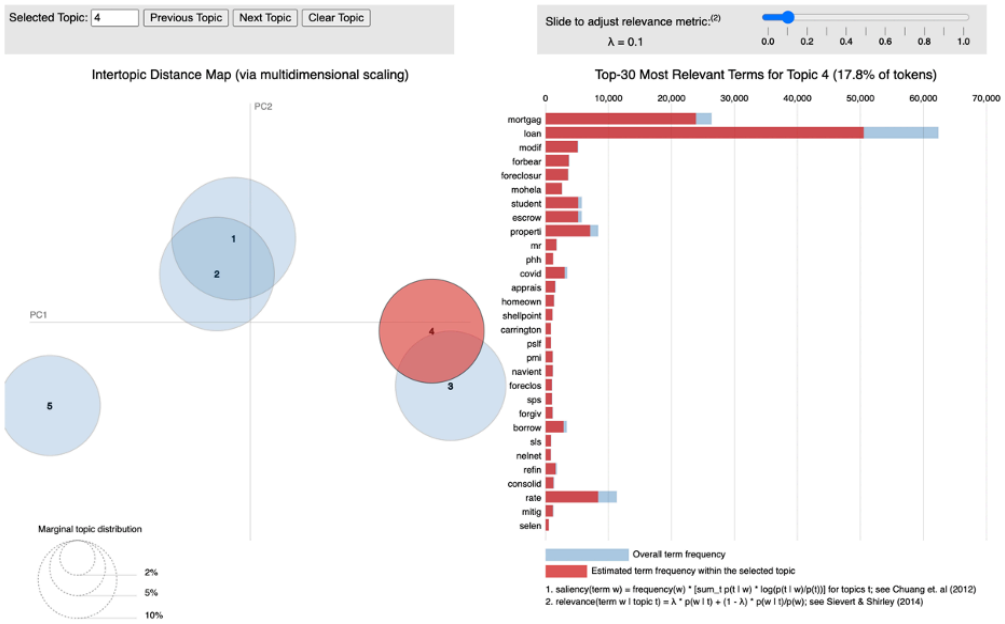
The first topic's prominent issues are reflected by common miscellaneous terms and jargon like "Recollect, Remove, Represent" prevalent in complaints. There are few product-specific words and financial terminology within this topic.



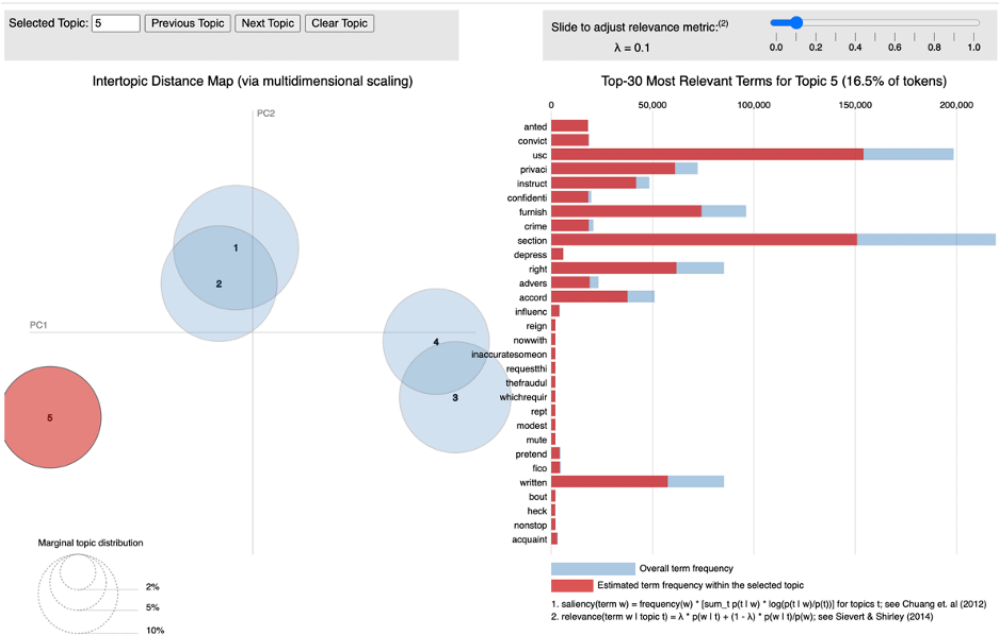
The second topic features regulatory terminology, indicating a focus on formal financial language within complaints. An inference can be made that regulatory jargon is a prominent textual trend within the complaints.



The third topic highlights money transfer issues related to institutions like Chase and Bank of America shown by words such as “Chase”, “Bank”, “America” and “Citibank”. “Wire”, “fund” and “transfer” all denote the strong emphasis on money transfers in this topic.



Topic four reveals concerns on mortgage obligations, including terms like “Foreclosure”, “Forebear” and “Appraise.” The terms are related to a specific sub sector of finance, namely real estate financing. Along with money transfers and specific institutions, real estate financing and the inability to meet debt payments is another significant topic mentioned within the complaint text.



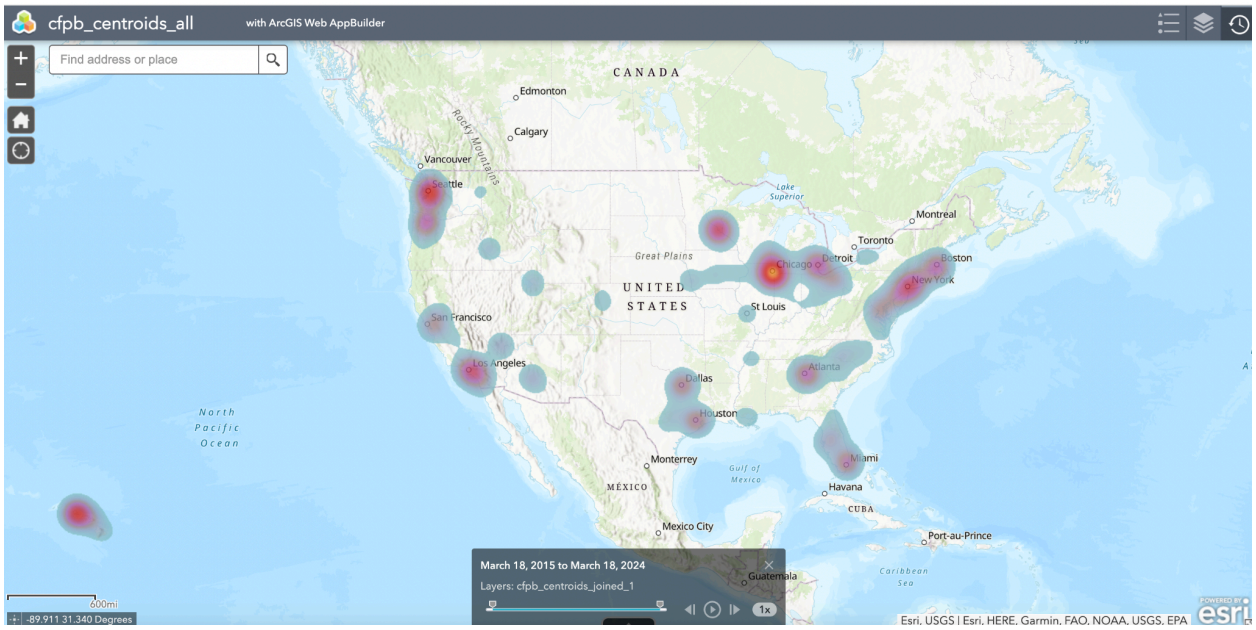
The final topic addresses privacy and data security, with terms like "confidential" and "FICO" being significant. The term “furnish” also alludes to furnishing one’s account, which is the process of banks sending consumer information to credit reporting agencies. A correlation can be drawn to the fact that complaints labeled as a credit reporting product issue is considered a significant majority within the dataset.

Word Cloud

The word cloud is constructed from terms extracted from complaint data that was received during the period of 2023 thus far. Terms related to credit and credit reporting are the predominant terms found within the corpus, possibly due to the significant number of cases involving credit and credit reporting. Other notable terms are related to bank account information and security issues.

ArcGIS Heat Map

The heat map embedded in our website shows the concentration of complaints by the mapped centroids via the zip code the complaints are sent from. The sliding time scale shows spatiotemporal changes in where the occurrences of complaints are densely located on an annual basis. We see that throughout 2015 to 2023, our analysis reveals a greater concentration of consumer complaints on the east coast, notably in urban areas like NYC, Chicago, and Atlanta, compared to the west coast, with significant activity in California.



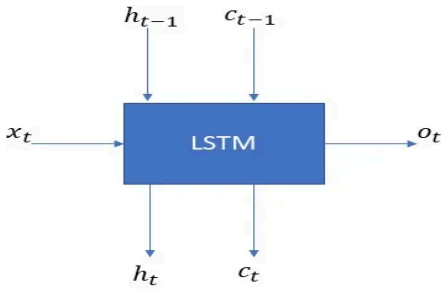
LSTM (Long Short-Term Memory)

We build a BiDirectional LSTM (Long Short-Term Memory) model with the TensorFlow module in a Python environment for multiclass text classification of complaints by their product types.

An LSTM model is a recurrent neural network model that can capture long-term dependencies within sequential data. Recurrent neural network models incorporate feedback connections rather than feedforward connections. Feedback neural networks are neural networks that allow for data to traverse through the signals in the network in both directions between the input state and the output state. This means that previous outputs from the LSTM model can affect new inputs. Feedforward models do not have this capability and only allow for data to travel in one direction from the input state to the output state. LSTM models address a drawback of Recurrent Neural Networks, which are vanishing gradients that make it difficult to measure long-term dependencies. A Bidirectional LSTM follows the same concept as a traditional LSTM model. However, it differs as it can process input data in both a forward-propagating and backward-propagating direction.

The vanishing gradient problem is defined as when the gradients of the cost function become smaller as the RNN model feedbacks from the output layer to the input layer. Recurrent neural networks utilize backpropagation, which updates the weights in each of the neurons using the partial derivatives of the cost function concerning weight. As shown from the chain rule, backpropagation becomes ineffective as a result of traversing backward through temporal intervals which makes it harder to update new weights onto pre-existing ones. The LSTM model shown below uses gates to control the flow of information and tackle the vanishing gradient problem. consists of a forget gate, an input gate, and an output gate. The forget gate takes into account both the previous hidden state (h) from T-1, and the current input X at t to filter what type of long-term information in the cell state is still relevant. The input gate also takes into account the previous hidden state and current input, to compute the magnitude and direction of how the elements in the cell state will be updated to represent the long-term information's relevance in a new way.

The output gate filters what information in the cell state is relevant. The newly updated cell state is sent to the network as input in the next period.

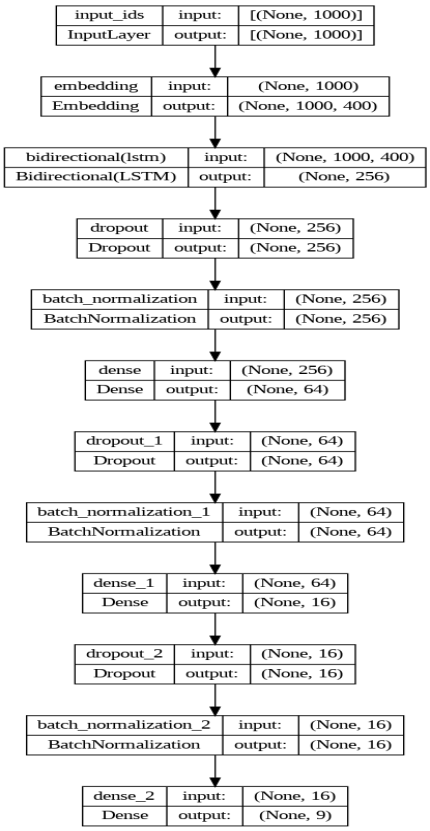


σ_g : $sigmoid$
σ_c : tanh
. : element wise multiplication

$$f_t = \sigma_g\,(W_f \times x_t + U_f \times h_{t-1} + b_f)$$
$$i_t = \sigma_g\,(W_i \times x_t + U_i \times h_{t-1} + b_i)$$
$$o_t = \sigma_g\,(W_o \times x_t + U_o \times h_{t-1} + b_o)$$
$$c'_t = \sigma_c\,(W_c \times x_t + U_c \times h_{t-1} + b_c)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t$$
$$h_t = o_t \cdot \sigma_c(c_t)$$

f_t *is the forget gate*
i_t *is the input gate*
o_t *is the output gate*
c_t *is the cell state*
h_t *is the hidden state*

We utilize the Keras API from the TensorFlow package, to construct a BiDirectional LSTM RNN suitable for sequential time series forecasting of a variable based on multivariate time series inputs. LSTM models require that we transform our sequential time series variables that will be used as predictors into a 3d array structured as (batch size, time window to use for prediction, and number of features). We shall also tune our hyperparameters (Learning rate, Beta 1, Beta 2 and Epsilon) of the TensorFlow LSTM model using the RandomSearch function, optimizing against the validation accuracy score. We intend to display the capabilities of Bidirectional LSTM models, for text processing. The CFPB only stores complaints of banks with more than $10 billion in assets. The potential utility that can be associated with this model, is that it can also be used within banks themselves to preemptively classify unstructured complaint data for faster processing and response times.




```
Epoch 1/8
441/441 [==============================] - 188s 412ms/step - loss: 0.6529 - accuracy: 0.7890 - val_loss: 0.5096 - val_accuracy: 0.8422
Epoch 2/8
441/441 [==============================] - 192s 437ms/step - loss: 0.5655 - accuracy: 0.8202 - val_loss: 0.4863 - val_accuracy: 0.8436
Epoch 3/8
441/441 [==============================] - 194s 439ms/step - loss: 0.5445 - accuracy: 0.8289 - val_loss: 0.4966 - val_accuracy: 0.8460
Epoch 4/8
441/441 [==============================] - 192s 435ms/step - loss: 0.5366 - accuracy: 0.8311 - val_loss: 0.5061 - val_accuracy: 0.8389
Epoch 5/8
441/441 [==============================] - 191s 434ms/step - loss: 0.5408 - accuracy: 0.8301 - val_loss: 0.5122 - val_accuracy: 0.8379
Epoch 6/8
441/441 [==============================] - 191s 432ms/step - loss: 0.5495 - accuracy: 0.8267 - val_loss: 0.5153 - val_accuracy: 0.8353
Epoch 7/8
441/441 [==============================] - 190s 431ms/step - loss: 0.5481 - accuracy: 0.8260 - val_loss: 0.4896 - val_accuracy: 0.8413
Epoch 8/8
441/441 [==============================] - 190s 430ms/step - loss: 0.5451 - accuracy: 0.8279 - val_loss: 0.4794 - val_accuracy: 0.8468
```

	precision	recall	f1-score	support
Credit reporting, credit repair services, or other personal consumer reports	0.63	0.57	0.60	10552
Debt collection	0.63	0.63	0.63	9575
Credit card or prepaid card	0.61	0.78	0.69	8868
Bank account or service	0.64	0.73	0.68	4106
Mortgage	0.00	0.00	0.00	2389
Money transfer, virtual currency, or money service	0.28	0.17	0.21	2308
Vehicle loan or lease	0.00	0.00	0.00	1260
Payday loan, title loan, personal loan, or advance loan	0.31	0.01	0.02	910
Student loan	0.93	0.96	0.94	100958
micro avg	0.84	0.84	0.84	140926
macro avg	0.45	0.43	0.42	140926
weighted avg	0.82	0.84	0.83	140926
samples avg	0.84	0.84	0.84	140926

Our results show that improvements can be made to increase accuracy scores within topics such as mortgage, money transfer and payday. Despite this, we see a weighted average f1-score of 83%. We can draw the conclusion that while overall performance of the multi-classification model is reasonably strong,

it is lacking in accuracy scores of specific product labels with smaller number of observations. Improvements can be made by potentially increasing the units within the hidden state of the BLSTM layer, as well as increasing the number of dense layers for extracting information of higher granularity.

Future Development

Expanding on the future development plan for our project in consumer complaint data analysis, our primary goal is to enhance the depth and breadth of our data sources. This involves not only regularly updating our database with the latest complaints and responses but also incorporating complementary datasets. By integrating broader economic indicators and consumer sentiment data, we can provide more nuanced and contextual insights.

In terms of analytical capabilities, we're looking to leverage advancements in Natural Language Processing (NLP) and sentiment analysis. These techniques will allow us to more accurately interpret the tone, urgency, and underlying issues in consumer complaints. In parallel, our predictive modeling will see significant improvements. We aim to explore and implement cutting-edge machine learning techniques, such as ensemble methods or deep learning networks, to improve the accuracy and efficiency of our predictions. Fine-tuning our existing LSTM models through advanced methods like cross-validation and hyperparameter optimization will further enhance their predictive power. A key component of our development plan is the implementation of real-time analysis. By creating a system that can ingest and analyze data as it's generated, we can offer immediate insights into current trends and issues, enhancing the responsiveness of financial institutions to consumer grievances. This will be complemented by an upgraded website interface that offers a more interactive and user-friendly experience, including customizable dashboards and live monitoring features.

Collaboration and knowledge sharing form another critical aspect of our plan. By engaging with financial experts, data scientists, and regulatory bodies, we can refine our analysis methods and ensure that our

findings are impactful and relevant. Sharing these insights with policymakers can assist in creating more consumer-friendly financial regulations and practices. Our scope expansion will see us venturing into international markets, conducting comparative studies to understand how consumer complaints and financial practices vary across different regions. This global perspective will enrich our analysis and provide a more comprehensive understanding of the financial landscape.Education and advocacy will be key priorities. We plan to publish detailed reports and white papers on our findings, and develop educational materials that empower consumers to better understand financial products and their rights. Furthermore, by providing tools for direct complaint reporting and tracking on our website, we aim to strengthen consumer advocacy and influence positive changes in financial services. To sustain and grow our project, we will seek collaborative partnerships with academic institutions for research opportunities and explore various funding sources. This will ensure the project's viability and its continued evolution in line with technological advancements and changing consumer needs.

Conclusion

This project provides a comprehensive analysis of consumer complaints within the financial industry. Our multifaceted analysis approach of using exploratory analysis, LDA topic modeling, and machine learning has yielded valuable insights on the dynamics of consumer grievances, issues, industry patterns and areas of improvement. We have uncovered various patterns in the types of issues consumers face and what areas can financial institutions improve to address consumer complaints.

Looking ahead, we focus on continuous improvement in data sources and analytical techniques. We plan to extend our analysis by incorporating international markets, collaborating with financial experts and broadening the scope of our research.

It is necessary to also acknowledge some of the limitations of our research. The analysis heavily relies on the CFPB dataset which only includes information related to larger financial institutions, therefore, in order to generalize our analysis results we must gain more relevant information from smaller financial institutions. As seen from the Heatmap of complaints across various states, there is a large discrepancy in the amount of complaints received from various states, making our sample of dataset biased. Finally, textual data always involves some sort of nuances and biases that cannot be fully captured through textual analysis.

In conclusion, our project focuses on utilizing a wide range of analytical tools to understand consumer concerns and behavior, while also pointing out how financial institutions can improve their products and services to keep their consumers satisfied.

References

[1] *Consumer Complaint Database*. (n.d.). Consumer Financial Protection Bureau. Retrieved December 18, 2023, from https://www.consumerfinance.gov/data-research/consumer-complaints/

[2] Oyewola, David & Lawal, Abdullahi & Julius, Sowore & Kachalla, Lummo & Dada, Emmanuel. (2023). Optimizing sentiment analysis of Nigerian 2023 presidential election using two-stage residual long short term memory. Heliyon. 9. e14836. 10.1016/j.heliyon.2023.e14836. https://www.researchgate.net/publication/369628149_Optimizing_sentiment_analysis_of_Nigerian_2023_presidential_election_using_two-stage_residual_long_short_term_memory/citation/download

[3] Pramod Kumar Naik, Prashanth T , S Chandru , S Jaganath , and Sandesh Balan . "Consumer Complaints Classification Using Machine Learning & Deep Learning." International Research Journal on Advanced Science Hub 05.05S (2023): 116–122. https://rspsciencehub.com/article_23795_3d7d1d4dfc63d072724eafdb92a5bcbf.pdf

[4] U.B. Mahadevaswamy, P. Swathi, Sentiment Analysis using Bidirectional LSTM Network, Procedia Computer Science, Volume 218, 2023, Pages 45-56, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.400

[5] Kaveh Bastani, Hamed Namavari, Jeffrey Shaffer, Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Expert Systems with Applications, Volume 127, 2019, Pages 256-271, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.03.001