# Bank Complaint Classification and Sentiment Analysis

Amratansh Sharma, Aditi Vatse, Prince Kumar, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

sharm321@purdue.edu; [avatse@purdue.edu](mailto:avatse@purdue.edu); kumar355@purdue.edu; lanhamm@purdue.edu

**ABSTRACT**

In this paper, we employ text mining and sentiment analysis on the complaints raised at Consumer Finance Protection Bureau portal to classify the complaint into 4 categories: Fraud, Fee, Policy, Unauthorized transactions. The purpose of our report is to demonstrate that the data can be mined to benefit both the financial institutions as well as CFPB in the following ways:

1. Financial institutions can use this to improve their response to the consumers and to identify the major concern areas which can be either in a specific product, in a specific region or most negative.
2. CFPB can use the analysis to identify the financial institutions which have a disproportionate share of the complaints and see if there is any trend in the complaint against a specific company or a specific product. Using this, they can preemptively investigate the issue before too many consumers are affected by it.

We have used R to achieve the goal. Both supervised and unsupervised learning were used. Techniques like sentiment analysis, text mining, classification tree and kappa score filtering.

Keywords: R, Text Mining, Sentiment Analysis, Supervised Learning, Unsupervised Learning, Classification Tree

**INTRODUCTION**

Consumer Financial Product Bureau (CFPB) was formed on July 21, 2011 with the vision to make consumer financial markets work for consumers, responsible providers, and the economy as a whole. They protect consumers from unfair, deceptive, or abusive practices and take action against companies that break the law. Since the formation of the bureau, consumers have raised more than 1.1 million complains on its portal about 13 types of financial products and services such as credit card, credit reporting, loans etc. The database makes and an excellent source for CFPB to monitor complaints against the financial industry and execute its regulatory responsibilities of market supervision and enforcement. At the same time, it is also an important tool for financial institutions to compare their performance against other companies in their field. In the monthly report published by CFPB, it provides a statistical summary of the consumers complaint data.

CFPB has investigated fraudulent practices by financial institutions which has led to many large fines in the last 2 years. Some of them are listed below:

1. In September 2016, Wells Fargo was fined a $190 million settlement with federal regulators and prosecutors for opening more than two million deposit and credit card accounts without customer authorization.
2. In March 2017, credit bureau Experian was fined 3 million dollars for selling misleading credit scores to their consumers.
3. In April 2108, Wells Fargo & Co. agreed on Friday to pay $1 billion in fines to the Consumer Financial Protection Bureau and the Office of the Comptroller of the Currency for harm inflicted on mortgage and auto loan borrowers.

Against this background, the objective of our study is to analyze the complaints and formulate a text analysis model using R to classify it in either of the four categories: Fraud, Fee, Policy, Unauthorized. Specifically, we are interested in exploring how financial institutions can utilize text mining for their internal audit function for early detection of fraudulent activities. Thus, the models we’ve developed can help financial services companies to predict business risks and potential issues based on the text data (customer complaints in this case). This can have a huge impact on the profitability of the financial institutions as the cost of any fraudulent activities is huge. With a better analysis of the complaints, the companies will be able to monitor such activities by any of their department and take action preemptively.

We are also designing a platform for financial institutions and CFPB where they can monitor trends in the complaint using various parameters available in the database. This would allow the organizations to see how they are performing against their competitors. It will also provide CFPB with an opportunity to compare complaint against a specific company to the industry.

In the next section, we present a brief review of papers which have conducted similar studies in text analytics using R and studies on the CFPB complaint database. In the next section we explain the methodology we have used for our analysis. Then we present the result of our analysis and make recommendations based on our study.

**LITERATURE REVIEW**

We went through multiple papers to understand the various methodologies used in text analytics, sentiment analytics. We also reviewed some papers which had earlier analyzed the CFPB database for various purpose. A summary of various papers is provided below:

Detecting Financial Fraud Using Data Mining Techniques

The aim of the paper was to review research studies conducted to detect financial fraud using data mining tools. As per their finding, logistics regression was the most prevalent technique used in the industry with a usage of 17%. Neural Network and decision trees were used by 15% of the organizations. Neural network which was primarily used in detecting credit rating fraud while decision trees was used in stock market prediction. Other techniques used were Support vector machines, Naïve Bayes, Bayesian, Discriminant analysis, K-Means etc. Authors have concluded that although logistic regression, decision tree, SVM, NN and Bayesian networks have been widely used (> 50%) to detect financial fraud, they are not always associated with the best classification results. Organizations may be able to select the most suitable technique once considering its usage context, frequency, and performance.

Detecting and Preventing Fraud with Data Analytics

The paper has advocated use of text analytics which includes text mining, sentiment analysis, text categorization etc. . The paper also advises to conduct geospatial analysis to understand the relevance of location where fraud happened so that we can determine any pattern in the behavior.

Integration of data analysis processes within the fraud detection system will help the company to get answers in real-time, creation of statistical analysis with high degree of accuracy increase the quality of analytical products. One of the challenges foreseen by the author is that due to the complexity of analytical research, the final product can be hard to assimilate, so, it’s recommended to use descriptive part (interpretation of tables, graphs, values, metadata etc.).

Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaint

This paper works on CFPB data source to create a model in SAS. The author uses SAS Contextual Analysis to find if the complaint requires any legal correction and customers greatest area of concern.

The paper uses SAS® Enterprise Guide to create SAS data set and the use contextual analysis. It provides details steps to find the trends in the customer narratives and the response to the customer complaints. Here they used a specialized sentiment model developed using SAS® Sentiment Analysis Studio® in its place. The results from this model were analyzed using visuals like decision trees. Finally, an interactive report to show the categorized complaints and responses by geography was created.

Using Text Mining and Natural Language Processing for Health Care Claims Processing

This paper is about processing healthcare claims, majorly related to insurance. Here, the author deals with both structured and unstructured data. Also, aims to bridge the gap between NLP and text mining. The model finally helps tag the claims as fraud and abuse, also, finds out if it related to specific insurer or not.

This problem is modeled using an innovative combination of NLP and semantic tags called NLP Concept Matcher. This process is both rule based, and statistic based natural processing using Content intelligence system. Here, first an exhaustive dictionary of the complete text is created. This is exhaustive as the data is limited to one domain. Once this is created, NLP is used to categorize the claims in broader categories of fraud or abuse.

Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study

The paper talks about applying text mining and sentiment analysis on the online data. It elaborates the process of text mining which includes converting the unstructured data into structured data which applies data mining techniques to do the preprocessing of data, extracts the patterns and relations from the text and visualize the patterns such as in the form of word clouds. Sentimental Analysis is also been applied in which they follow the lexicon based sentiment analysis approach. The main proposed methodology which they apply over here is data access, data cleaning, data analysis and visualization.

From Words to Actions: Using Text Analytics to Drive Business Decisions

The data chosen for the analysis comes from the Wells Fargo complaints. There were 4,643 complaints that included a consumer narrative and of those 756 resulted in the bank providing some sort of relief. They applied unsupervised modeling in which they performed text topics generation. The strategy went with Data preparation, in which they did topic clustering and created a text topic node. Also, they did apply supervised learning in which they focused on predicting avoidable complaints. It led them create the rule builder node, which attempts to predict which complaints received relief by creating a variety of AND/AND NOT combinations of individual terms and using them as predictors. Also, they created a text profile node which provides corroboration of these results using a hierarchical Bayesian model to predict which terms are the most likely to accurately distinguish whether a customer received relief (1 or 0).

Analyzing customer experience feedback using text mining: a linguistics-based approach

The paper suggests a linguistic based text mining model to inform the process of developing an improved framework. This framework incorporates important element of customer experience by encompassing three value creation elements: Activities, Resources, Context (ARC framework)

The paper suggests improvement in the following aspects:

1. Analyzing customer service at different periods to analyze the impact of specific service change
2. Customer feedback should be generated not just one factor but multiple improvement factors

The paper can be further replicated for other domains and industries. We can also investigate how the results of text mining can be included in the information systems of organizations so that maximum value is derived out of it.

Skeletons in the Database: An early analysis of the CFPB’s Consumer Complaints

The paper analyzes a database of approximately 110,000 consumer complaints lodged with CFPB and tries to identify various patterns that can be observed. Some of the outputs were:

1. Two major banks BofA and Citi were significantly less timely than average in responding to customer request.
2. Consumers of larger financial institutions were significantly more likely than the average consumer to dispute the initial response to their complaints

The paper also showed that regression analysis suggests that consumer financial companies respond differently to complaints, depending on the type of product and issues involved, thereby generating significant differences in the timeliness of responses and whether consumers dispute those responses. The paper uses various demographic factors such as income, education, race, unemployment etc. to analyze the data.

Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach

This study examines the impact of various features online user reviews such as basic, stylistic and semantic characteristics and tries to determine if they help in getting a higher “usefulness: rating from other users. They have applied text mining techniques to extract semantic characteristics from reviews. Two most important factors determine the usefulness of the reviews are: semantics of the text and how extreme the views are. Reviews with extreme opinion receive more helpfulness votes than ones with neutral or mixed opinions. They have used LSA based text mining technology which includes text preprocessing, parsing, term reduction, singular value decomposition and factor analysis.

Sentiment analysis algorithms and applications : A survey

This survey paper performs a comprehensive review of various sentimental analysis algorithms and the latest update in the field. They have broadly divided sentimental classification techniques into Machine Learning approach and lexicon-based approach. Machine Learning approach is further divided based on types of classifier: Decision Tree Classifier, Linear Classifier which includes Support Vector Machines and Neural Network, Rules based classifier and probabilistic classifiers which include Naïve Bayes, Bayesian Network and Maximum Entropy.

Lexicon based approach consists of both dictionary-based approach and corpus based approach which is further divided into statistical and semantic techniques. Naı¨ve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem. They are considered a reference model where many proposed algorithms are compared to.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Name of the Paper** | **Initiative and findings** | **Incites Obtained** |
| 1. | Detecting Financial Fraud Using Data Mining Techniques | Used decision trees, logistic regression, neural networks, SVM to find the financial fraud. | The algorithms used in the paper although good for classification but not always gives the best results. It depends on usage context, frequency and problem type |
| 2. | Detecting and Preventing Fraud with Data Analytics | Elucidates the use of text analytics and, explains about the text mining, sentiment analysis and conduct \geospatial analysis. | How different text analytics techniques can be integrated with each other to perform fraud detection with high accuracies. |
| 3. | Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaint | Create a model in SAS, having data source as CFPB. The author comes out with SAS Contextual Analysis to find customers area of concern. | How contextual analysis can be used to find a trend in the customers complaints and narratives, also how responses can be seen by geography. |
| 4. | Using Text Mining and Natural Language Processing for Health Care Claims Processing | Specifically the model was designed for the healthcare claims, related to insurance. Uses NLP to categorize the claims in broader categories. | How NLP and text mining are connected. The usage of NPL Concept matcher which can be another way of tagging the complaints. |
| 5. | Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study | The author mainly focuses on application of text mining, sentiment analysis on the online data and how information can be extracted from the unstructured texts. | How unstructured data can be converted into the useful patterns and trends in the information using data access, cleaning, analysis and visualization. |
| 6. | From Words to Actions: Using Text Analytics to Drive Business Decisions | The study was related to the Wells Fargo case. The author applied unsupervised learning to come out with text topic generation. Also they applied supervised learning for the avoidable complaints. | How supervised and unsupervised learning can be performed on the parts of complaints to come out with the meaningful results and the application of Bayesian Model to judge the satisfaction of customers. This process we thought, is the post work which which we are doing now. That is first classifying the complaints and then classifying the responses to the customers. |
| 7. | Analyzing customer experience feedback using text mining: a linguistics-based approach | The paper talks about the linguistic based approach to come out with the improved framework. Also, the effect of recording the customers behavior on a timely manner and the factors involved in this. | The framework of Activities, Resources and Context(ARC) and how text mining can be involved in the information systems to come out with the awesome values. |
| 8. | Skeletons in the Database: An early analysis of the CFPB’s Consumer Complaints | Uses the CFPB as a data source to identify various patterns. It comes out with the differences between the financial institutions.  They analyze the customer behavior to change there methodologies and also analyze the company’s response based on regression. | How results from the technique such as regression can be varied depending on the type o issues and conditions and how the different company’s performance can be varied based upon the analysis of customers behaviors trends and response. Basically, the goal is to come out with incites which makes an impact |
| 9. | Exploring determinants of voting for the “helpfulness” of online user reviews: A text  mining approach | The author basically deals with the online user reviews to measure their work and how they can improve for getting a higher user rating. The application of LDA based text mining helps in finding the extreme in opinions by the proper pre processing of data. | The LDA Based approach and the preprocessing to extract out the incites from the reviews and how the company can improve the performance of products using proper pre processing of the reviews using text mining. |
| 10. | Sentiment analysis algorithms and applications:  A survey | The survey basically talks about the sentimental analysis and the classification of sentimental analysis into machine learning approach and lexicon based approach. | The difference in algorithms which falls under the machine learning and lexicon based technique. From here we got the idea to go for the classification tree. |

**DATA**

We have downloaded data from the CFPB database available at <https://data.consumerfinance.gov/api/views/s6ew-h6mp/rows.csv?accessType=DOWNLOAD>

The data consists of the following rows:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Date received | Date | Date of complaint registration |
| Product | Categorical | One of the options provided in Annexure |
| Sub-product | String | Description provided by consumer |
| Issue | String | Description provided by consumer |
| Sub-issue | String | Description provided by consumer |
| Consumer complaint narrative | String | Description provided by consumer |
| Company public response | Categorical | One of the options provided in Annexure |
| Company | Categorical | Name of the company |
| State | Categorical | State of the consumer |
| ZIP code | Categorical | Zip Code of consumer residence |
| Tags | String | Description provided by consumer |
| Consumer consent provided? | Categorical | One of the options provided in Annexure |
| Submitted via | Categorical | One of the options provided in Annexure |
| Date sent to company | Date | Date |
| Company response to consumer | Categorical | One of the options provided in Annexure |
| Timely response? | Categorical | Yes/No |
| Consumer disputed? | Categorical | Yes/No |
| Complaint ID | Unique ID | Unique ID of each complaint |

**METHODOLOGY**

We followed the basic data mining flow, wherein first worked on data understanding and processing and then created model using the same.

Acquiring train data

First, we started will creating train data to classify the complained. This was done by tagging randomly selected 4,532 complaints to divide them in 4 categories : frauds, unauthorized transaction, misleading information or policy, and penalty fee. These 4,532 complaints serves as train data for complete model. All the columns from train data are removed except complaint ID and complaint text columns.

To generate the tags, we invited 137 volunteers to help us tagging. We developed a structured tagging protocol that provided examples of how to tag consistently among all tagging volunteers. Next, we distributed this protocol to the volunteers, had them read the instructions and begin classifying each complaint they received based on four categories. On average, each tagger tagged 150 complaints and each complaint was tagged by 4 individuals.

In addition, authors also tagged 360 complaints. These tags were treated as the reference for calculating kappa score in the next step.

The tagging process created a dataset with 7 columns:

1. Complaint ID and corresponding complaints
2. Tagger ID (email address)
3. Four target variables.

These tagged complaints were distributed in 4 separate files, one for each category for further processing. Below process was repeated for each file.

Validating

The dataset was split into two different datasets, “protag” which is tagged by authors and “regular” which were tagged by our volunteers. We compare the performance of each student against the tags by the authors on the overlapping dataset. The kappa score indicates whether the taggers were doing their work correctly. Tags from taggers with more than 0.4 were kept. The dataset was latter split into four datasets, each contains Complaint ID and corresponding complaints, tagger ID and corresponding kappa score, and one single target variable.

Finalizing

The final target variable for each of the four unique topics was based on two different methods, voting and kappa score oriented.

Voting method means that if two or more people voted “Yes” on a topic, that complaint topic was assigned as a “Yes,” otherwise it was assigned as a “No”. Kappa score oriented method determines the complaint topic based on the taggers that has the highest kappa score.

Out of the final eight datasets were generated, only one file for each type of category is chosen. Each file only contains the complaint content and one target variable.

Modeling

Once we had these datasets, we divided the data in 2 parts: Train (75%) and Test (25%). Train was used to train a predictive model wherein we would fit in the complaint and it would provide a result confirming if it belonged to a specific category or not.

Text Mining

Our initial approach was to tokenize each of the complaints using ‘tidytext’ library from R and create a data dictionary that would uniquely identify each of these categories. But once we created a frequency map of the most common 500 words appearing in each type of the complaint, we found that 340 of those words were in common. Removing those words would not have made sense.

Hence, we decided to use the classification tree algorithm ‘rpart’ to predict the category of the the observations. Once we had the model, we ran the model n test data and checked the accuracy, results of which are summarized in the results section. The purpose of the classification trees is to learn how we can discriminate between the various categories based on the tokens of each complaint. Tree classification techniques produce accurate predictions or predicted classifications based on few logical if-then conditions, have several advantages over many of those alternative techniques the most important of which is simplicity of results. This simplicity is useful not only for purposes of rapid classification of new observations but can also often yield a much simpler "model" for explaining why observations are classified or predicted in a manner.

Sentiment Analysis

We also conducted a sentiment analysis using the ‘nrc’ library in R and visualize what kind of sentiments are present in the maximum number in all the complaints. Sentiment Analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Here we have done sentiment analysis on the complaints. We use Qdap library which gives us the polarity score and also if we want to know that what number of positive and negative words are present, the function also provides us with these stuffs.

**MODEL ANALYSIS**

Classification Tree:

Classification tree is similar to the regression tree, the only difference is that it is used for predicting qualitative responses. In classification tree, we classify each observation belongs to the most commonly occuring class of the training observations. We are not only interested in the class prediction but also in the proportion of the class among training observations that falls into that region. Here criterion for the binary split is classification error rate. Gini index and entropy are the two other criterion which are used for the tree growing. Sometimes it happens that some of the splits yield two terminal nodes that have same predicted values but still split is performed because it increases the node purity. The accuracy of the model we applied changes by changing the frequent terms we take and the sparsing we do.

Multinomial Logistic Regression:

Multinomial Logistic Regression is the classification technique which is used for generalizing the logistic regression to the multiclass problem. having more than two discrete outcomes. It is basically used when the dependent variable in the problem is nominal. We tried multinomial logistic regression but the drawback we found was that it does not run when we do have very large number of predictors. In our model, the predictors are actually the word which we have got after the pre processing and as they are in large numbers it generally throws an error complaining about the size of predictors.

Random Forest:

Random forest has got a very nice overcome on the drawback of bagging. In bagging the subsets are randomly chosen and the average is taken after that again and again to reduce the variance. But suppose if there is some very strong predictor then, every subset will take that variable and then most of the sample trees would look same. However, random forest overcome this drawback by selecting some m randomly chosen predictors out of p predictors in every model and make sure that in every split it considers different m set of predictors. This way the correlation decreases. Also, this is called decorrelating the trees. When we apply random forest in our model, it really takes a lot of time, some hour or so and hence we decided to switch over to the classification tree.

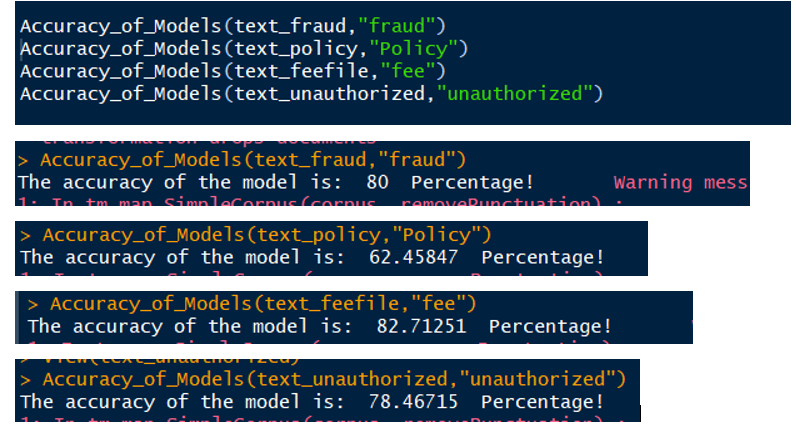
**RESULT**

Our model had different level of accuracies for each of the categories:

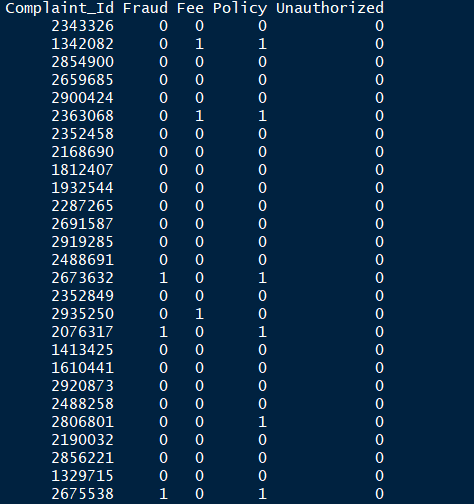
|  |  |  |
| --- | --- | --- |
| S No. | Category | Accuracy % |
| 1 | Fee | 82.7 |
| 2 | Fraud | 80.0 |
| 3 | Unauthorized | 78.5 |
| 4 | Policy | 62.5 |

The original complaint goes through all the four models and then we finally combine all the four data frame with the complaint ID. Also, we found that the fraud complaints are the highest in number and the unauthorized complaints are smallest in number. This also makes sense because when we observe the polarity of every category of complaint we found that fraud category has got the highest negative polarity among all the four. So, the company should look into more number of the fraud cases to sort it out and reduce the number of complaints in future.

Complaint Classification Results:

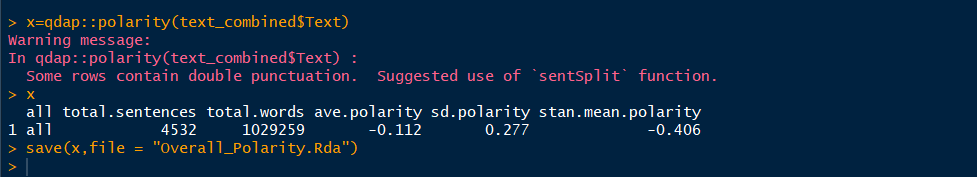


Output File Structure:

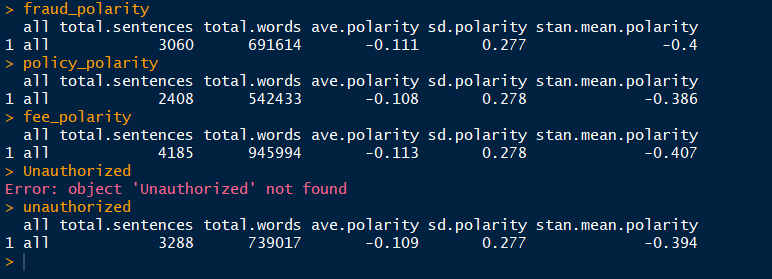


Sentiment Analysis Results:

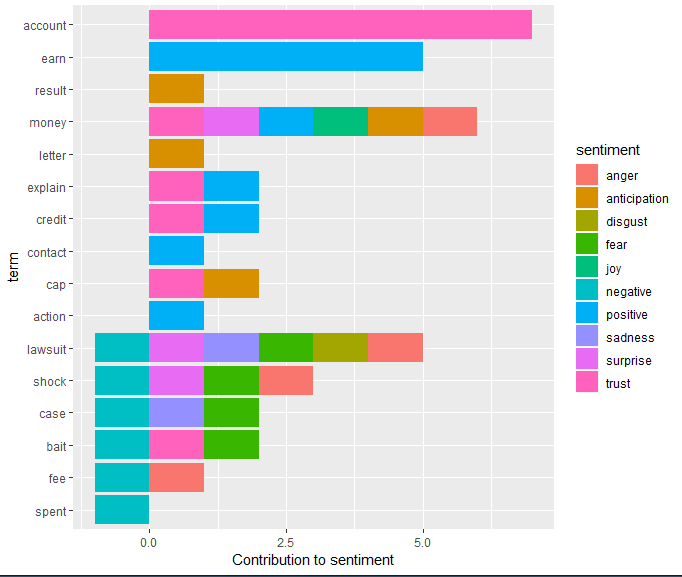
Overall Polarity



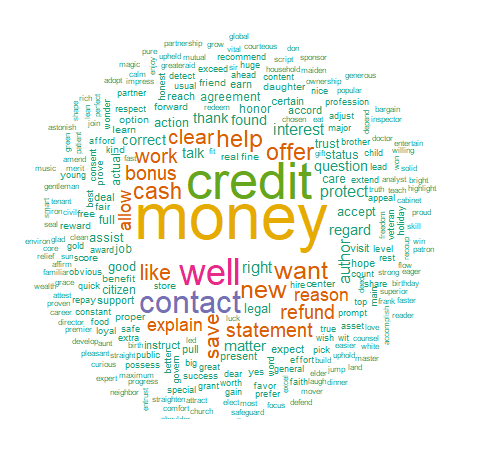
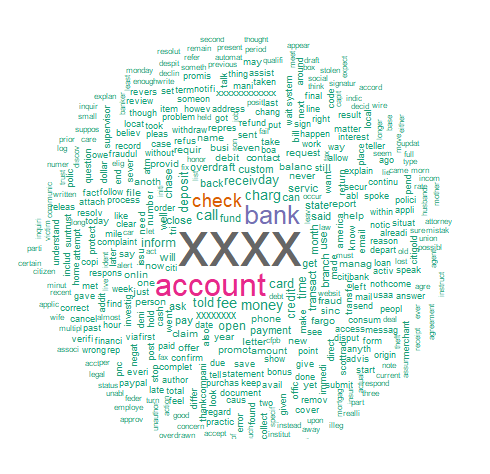
Individual Complaint Type Polarity



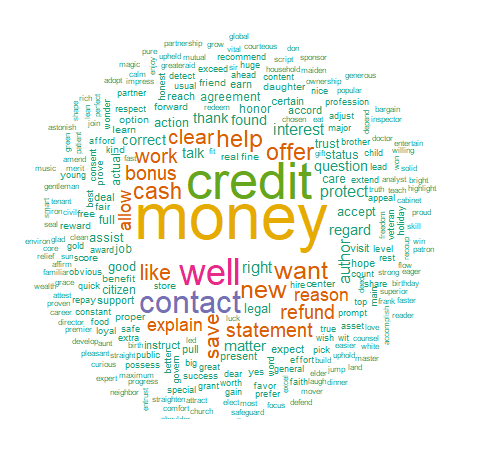
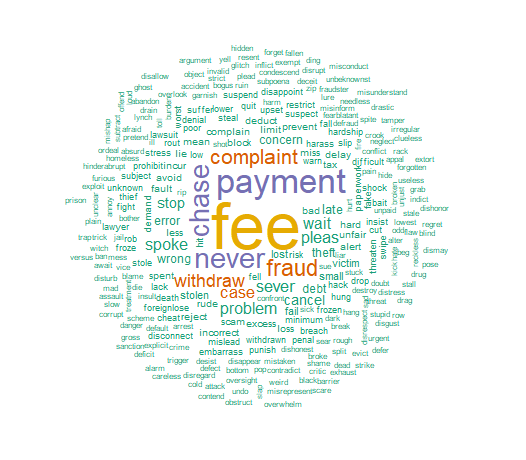
Sentiment Contribution



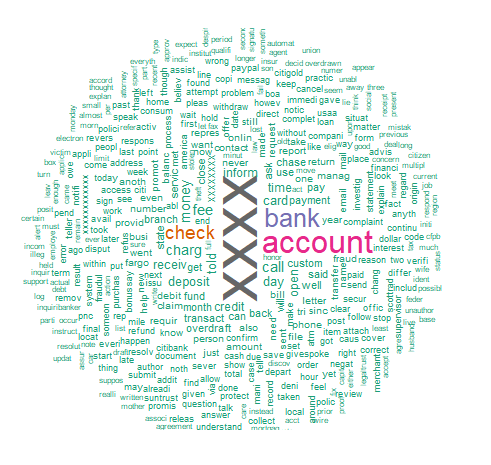
**Word Cloud**:



**Fig 1: All Complaints** **Fig 2: Fraud Complaints**

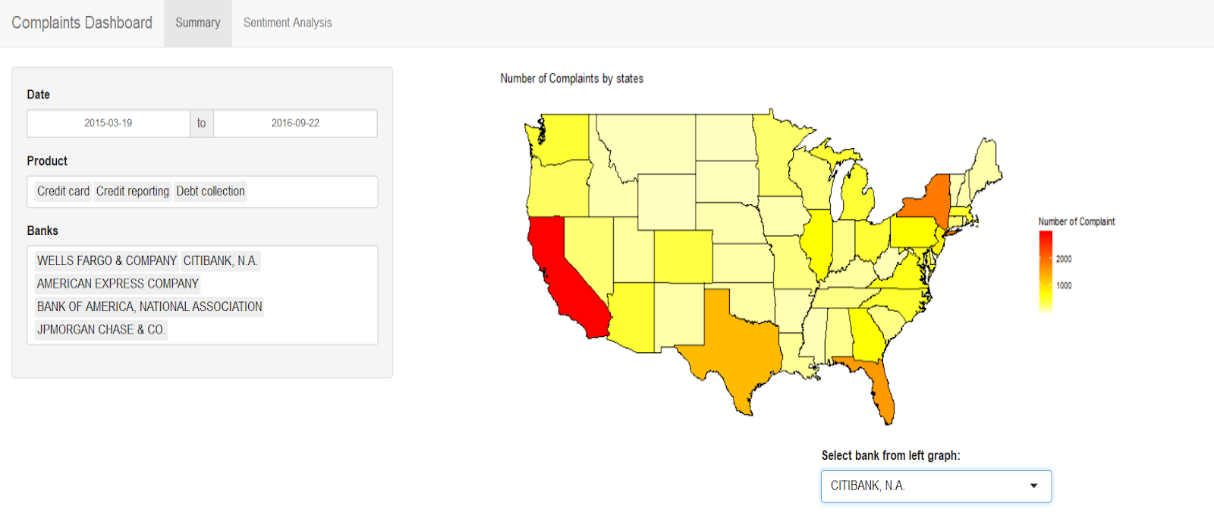


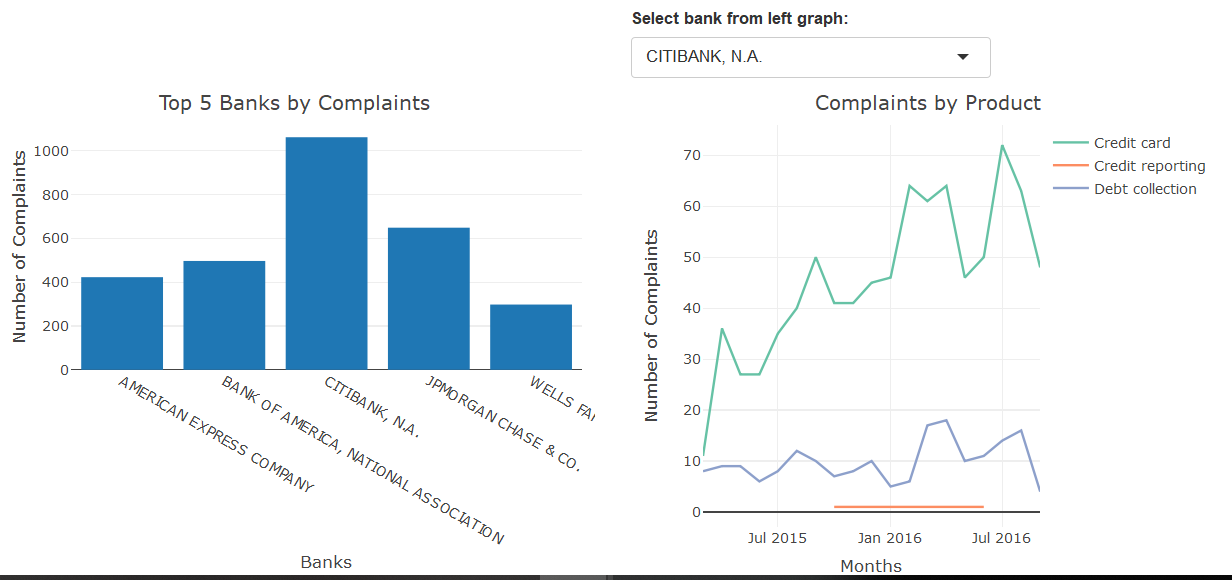
**Fig 3: Fee Complaints**  **Fig 4: Policy Complaints**

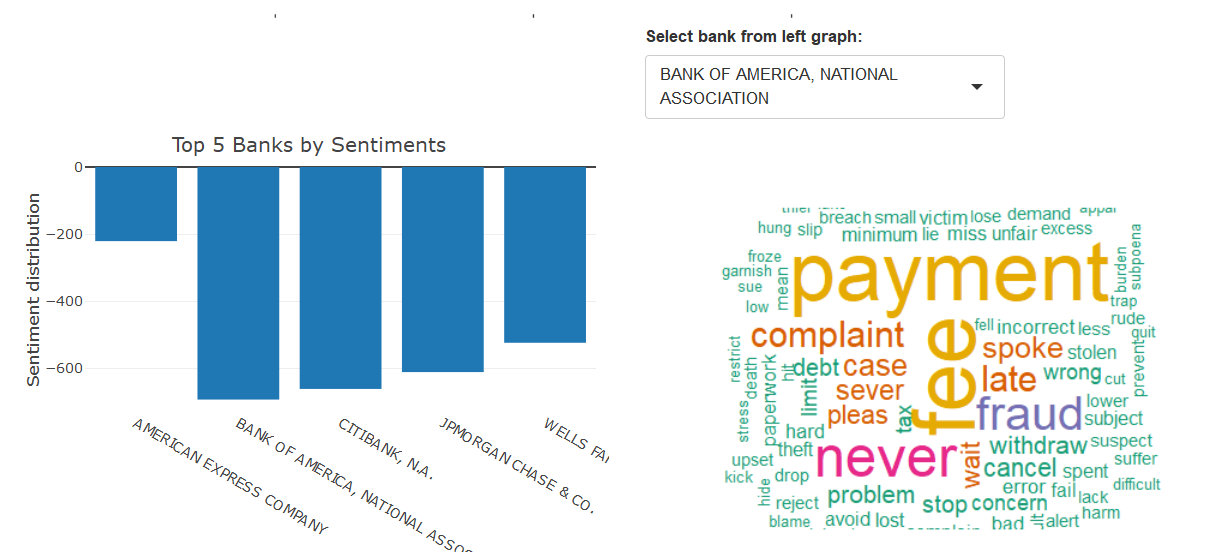


**Fig 5: Unauthorized Complaints**

**Proposed App screenshots**:







**CONCLUSIONS**

Fraudulent cases from huge banks is something that cannot be ignored in an era where data is so easily available. Hence, with our project, we not only aim to provide the classification and sentiments but provide greater visibility for these output. Since, we have huge dataset, best way to draw insights is through providing an app to CFPB personnel or even, the banks for internal audit. We have created a mock version of such an app to reason our final solution [Appendix: Proposed App]. The current limitations of our model is four separate columns for each category of complaint. This could be combined into one and used as part of the app. Also, we can aim to have customized section for banks where in details of stakeholders in each state for each complaint could help have a transparent complaint processing system. Another limitation is that in the current research, we have not considered the number of consumers each institution has. If we use that information, we can compare the performance of institutions in a more accurate way.

**REFERENCES**

1. <https://data.consumerfinance.gov>
2. <http://journals.sagepub.com/doi/abs/10.1177/1094670514524625?journalCode=jsra>
3. <http://www.jds-online.com/file_download/558/%E6%94%B910-Detecting+Financial+Fraud+Using+Data+Mining+Techniques-JDS_V3.pdf>
4. <http://journals.sagepub.com/doi/abs/10.1177/1094670514524625?journalCode=jsra>
5. <https://analytics.ncsu.edu/sesug/2017/SESUG2017_Paper-202_Final_PDF.pdf>
6. <https://pdfs.semanticscholar.org/4db5/cc991d9471b75e6fdd015531654b4a9819bd.pdf>
7. <https://dl.acm.org/citation.cfm?id=1089824>
8. <http://support.sas.com/resources/papers/proceedings17/SAS0282-2017.pdf>
9. <http://digitalcommons.law.yale.edu/fss_papers/4876/>
10. <https://www.cnbc.com/2016/09/08/wells-fargo-reaches-185m-settlement-to-settle-secret-account-fraud-case.html>
11. <https://www.consumerfinance.gov/about-us/newsroom/cfpb-fines-experian-3-million-deceiving-consumers-marketing-credit-scores/>
12. [https://www.npr.org/sections/thetwo-way/2018/04/20/604279604/wells-fargo-hit-with-1-billion-in-fines-over-consumer-abuses\](https://www.npr.org/sections/thetwo-way/2018/04/20/604279604/wells-fargo-hit-with-1-billion-in-fines-over-consumer-abuses%5C)
13. <https://www.acl.com/pdfs/ACL_fraud_ebook.pdf>
14. <https://www.sciencedirect.com/science/article/pii/S0167923610001909>
15. <https://www.sciencedirect.com/science/article/pii/S2090447914000550>
16. Report from last work on CFPB database by students at Purdue University

**APPENDIX**

Categorical Responses in Data

|  |  |
| --- | --- |
| Product | Consumer Consent Provided |
| Bank account or service | Consent not provided |
| Checking or savings account | Consent provided |
| Consumer Loan | Consent withdrawn |
| Credit card | Submitted via |
| Credit card or prepaid card | Email |
| Credit reporting | Fax |
| Credit reporting | Phone |
| Debt collection | Postal mail |
| Money transfer, virtual currency | Referral |
| Money transfers | Web |
| Mortgage | Company Resonse to Customer |
| Other financial service | Closed |
| Payday loan | Closed with explanation |
| Payday loan, personal loan | Closed with monetary relief |
| Prepaid card | Closed with non-monetary relief |
| Student loan | Closed with relief |
| Vehicle loan or lease | Closed without relief |
| Virtual currency | In progress |
|  | Untimely response |