Identification of Nature of Complaint from Consumer Complaint Database using NLP

Sayanka Mukherjee

Under the Guidance of **Prof. Faiz Ahamad**

Abstract: ***Financial products can assist us in increasing the amount of money we have in order to achieve various financial objectives, such as retirement, children's education, marriage, and so on. But sometimes the financial products are not handled in ways they should be handled and that is where a customer faces a crisis. Often financial organizations create a relationship or a contract over financial products delivered to a customer. Consumer Complaint database is once such pool of complaints lodged by customers from various corner of the country facing issues over these financial products. The Consumer Complaint Database is a collection of complaints about consumer financial products and services that we sent to companies for response. Hence, I propose a TF-IDF based Information retrieval model that would make the classification of complaints made by consumer much less time consuming and faster. The goal is making a computer capable of understanding the contents of documents, including the contextual nuances of the language within them. This then is used to extract information and insights contained in the documents as well as categorize and organize the documents themselves.***

Keywords: Term Frequency, Inverse Document Frequency, Vectorization, Natural Language Processing.

1. Introduction

The Consumer Complaint Database is a collection of complaints about consumer financial products and services that we sent to companies for response. Complaints are published after the company responds, confirming a commercial relationship with the consumer, or after 15 days, whichever comes first. Complaints referred to other regulators, such as complaints about depository institutions with less than $10 billion in assets, are not published in the Consumer Complaint Database.

The database generally updates daily. Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint. Consumers must opt-in to share their narrative. Govt. will not publish the narrative unless the consumer consents and consumers can opt-out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer. Companies' public-facing responses to complaints are included if companies choose to publish one. Companies may select a public response from a set list of options as soon as they respond to the complaint, but no later than 180 days after the complaint was sent to the company for response, e.g. "Company believes complaint is the result of an isolated error".

Complaints can give us insights into problems people are experiencing in the marketplace and help us regulate consumer financial products and services under existing federal consumer financial laws, enforce those laws judiciously, and educate and empower consumers to make informed financial decisions.

But gathering such a large pool of complaints and classifying them into correct subclass of complaints is a very manually expensive and time-consuming process. Complaint volume should be considered in the context of company size and/or market share. For example, companies with more customers may have more complaints than companies with fewer customers.

The sentiments and the ontological value of text is often not given the importance it deserves and this is where Natural language processing plays a very vital role. NLP is concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is making a computer capable of understanding the contents of documents, including the contextual nuances of the language within them. This then is used to extract information and insights contained in the documents as well as categorize and organize the documents themselves. TF-IDF is one such statistic measure of NLP that is able to capture the context and sentiment of the corpus very well. The TF-IDF algorithm is followed by a Machine Learning algorithm that will classify the complaint based on its content.

This is a supervised machine learning classification task as it will be performed on a labeled dataset. The model will be trained on past historic data where the nature of each complaint is manually identified and then classified. Classification algorithms like Linear Support Vector Machine (LinearSVM), Random Forest, Extreme Gradient Boosting, Multinomial Naive Bayes and Logistic Regression will be applied on vectorized data and the performance of all these models will be compared. The metrics of comparison of these models will be Accuracy, Precision, Recall and F1-score. Then a confusion matric based comparison will be made to find out the True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate. Finally, Classification Reports of all models will be generated to compare the number of correctly classified complaints of each class.

2. Objective

The main research objective through the proposal of this model is, to build a model that would make the process identification of type of complaint lodged by the customer about a financial product or service, faster, precise and more accurate. The next objective is to capture the real sentiment of the customer complaint so that companies can nurture better ways to solve the issues occurring withing their services. The last objective is to realize the most affected financial product/service among the given preprocessed whole updated data.

The main questions that the proposed research wishes to answer are: Can this model understand the context of the complaint and capture the real sentiment of the corpus? Can the true positive rate be increased false positive rate be decreased by optimizing the model? Can this model having been transferred in cross industry and cross dialect field of research give similar results as in English?

3. Literature Gap

We have referred to multiple articles that have used Data, Statistics, machine Learning, deep learning etc. like technologies to solve finance related real life problems, but the customer satisfaction region has been untouched by most of them.

Carsten and Lennart have used Data Analysis and Big data as an important tool to analyze survey of bank customers to realize that most customers wanted introduction of new technologies into banking process to make it faster and simpler. Within the framework of inductive statistics, the statistical software "R" was used to refute or prove hypotheses with statistical tests. These statistical tests were carried out with a significance level of $\alpha = 0.05$. Six statements were hypothesized and proved as important or unimportant with respect to the context by using Exact Binomial test. But it seems to miss out on whether these customers like a particular product or service that the bank offers.

Stefan and Jonas use sentiment analysis of public tweets ad social messages to help determine stock movements. They use NLP to assign these social messages to their respective stock names. Also, they use language filtering and POS tagging to understand the semantic nature of the text. Then they work on finding the nature of messages as positive or negative on basis of selling or buying of stocks. A similar NLP based text processing technique can be used on our approach where we can use the complaints to understand their nature of speech using POS tagging and also filtering out filthy and nasty words without changing the meaning the complaint wants to convey.

Nikolas and Paolo improve the credit risk management procedure by using machine learning models to extract non-linear relations among the financial information contained in the balance sheets. They use the Extreme Gradient Boosting algorithm because it improves tree models strengthening their classification performance. They improve the choice selecting models based on their predictive accuracy, and employing a posteriori an algorithm that achieves explanability. We will also use the XGB model for the classification task as it will help us achieve better performance metrics by improving on the relations of the model parameters in POS tagging.

4. Methodology

The dataset contains different information of complaints that customers have made about a multiple products and services in the financial sector, such us Credit Reports, Student Loans, Money Transfer, etc.

The date of each complaint ranges from November 2011 to May 2019. This work is considered a U.S. Government Work. The dataset is public dataset and it was downloaded from the website https://catalog.data.gov/dataset/consumercomplaint-database. The given dataset contains 1282355 complaints and 17 features on which the target column is decided. The dataset contains features that are not necessary to solve our multi-classification problem. For this text classification problem, we are going to build another dataframe that contains ‘Product’ and ‘Consumer complaint narrative’ (renamed as 'Consumer_complaint'). On removing all the null values from the columns we get a final dataframe of 3,83,584 complaints.

There are 18 different classes or categories (target). However; it is observed that some classes are contained in others. For instance, ‘Credit card’ and ‘Prepaid card’ are contained in ‘Credit card or prepaid card’ category. Now, imagine there is a new complaint about Credit card and we want to classify it. The algorithm can either classify this complaint as 'Credit card' or 'Credit card or prepaid' and it would be correct. Nevertheless, this would affect model performance. In order to avoid this problem, the names of some categories were renamed.

The bar chart below shows the number of complaints per category. It can be observed that The bar chart 4.1 below shows the number of complaints per category and 4.2 shows percentage complaints for each classes. It can be observed that most of customer complaints are due to: credit reporting, credit repair, debt collection and mortgage.

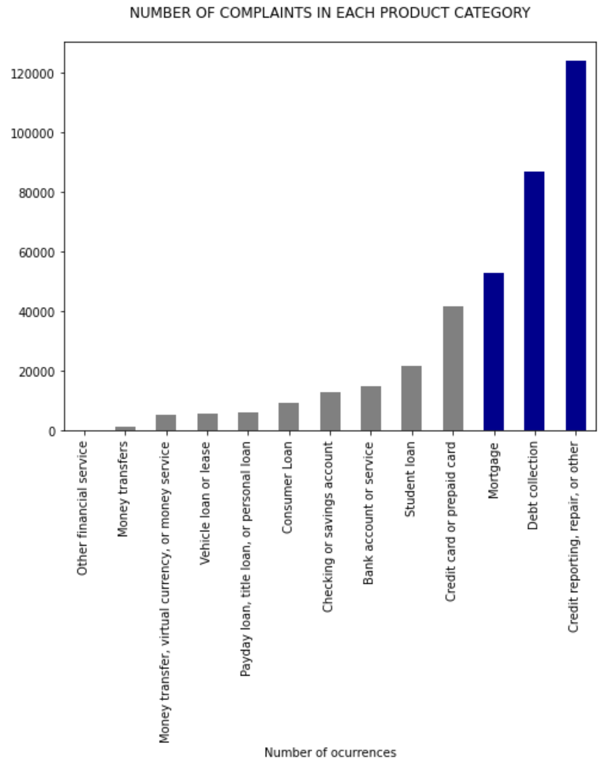

Chart 4.1: Count of complaints of each category

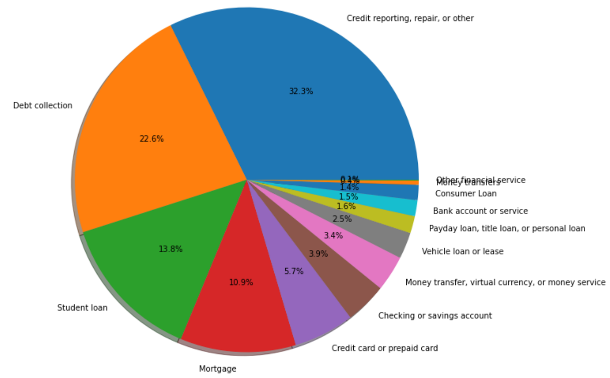

Chart 4.2: Percentage of classes of complaints

The text needs to be transformed to vectors so as the algorithms will be able make predictions. In this case it will be used the Term Frequency – Inverse Document Frequency (TFIDF) weight to evaluate how important a word is to a document in a collection of documents. After removing the punctuation and lower casing the words, importance of a word is determined in terms of its frequency. Term Frequency – Inverse Document Frequency TF-IDF is the product of the TF and IDF scores of the term.

Term Frequency summarizes how often a given word appears within a document. Inverse Document Frequency downscales words that appear a lot across documents. A term has a high IDF score if it appears in a few documents. Conversely, if the term is very common among documents (i.e., “the”, “a”, “is”), the term would have a low IDF score.

TF-IDF are word frequency scores that try to highlight words that are more interesting, e.g. frequent in a document but not across documents. The higher the TFIDF score, the rarer the term is. For instance, in a Mortgage complaint the word *mortgage* would be mentioned fairly often.

However, if we look at other complaints, *mortgage* probably would not show up in many of them. We can infer that *mortgage* is most probably an important word in Mortgage complaints as compared to the other products. Therefore, *mortgage* would have a high TF-IDF score for Mortgage complaints.

5. Results / Discussion

The TFIDF vectorizer produces the nearest/closest bigrams and trigrams to the class/target names. Table 5.1 shows the nearest n-grams.

Table 5.1: Classes v/s bigrams v/s trigram

Class	Bigram	Trigram
Bank account	Overdraft bank	Debit card checking
Savings Account	Checking branch	Overdraft fees bonus
Consumer loan	Vehicle dealership	Car loan acceptance
Credit card	Citi card	American express balance
.Credit reporting	Experian equifax	Credit file report
Debt collection	Collection amount	Collection agency report
Money Transfer	Coinbase ethereum	Account coinbase bitcoin
Mortgage	Escrow modification	Short sale company
Payday loan	Astra payday	Applied payday loan
Student loan	Navient student	Income based student
Vehicle loan	Honda car	Used vehicle finance

The original data was divided into features (X) and target (y), which were then splitted into train (75%) and test (25%) sets. Thus, the algorithms would be trained on one set of data and tested out on a completely different set of data (not seen before by the algorithm). The classification models evaluated are: Random Forest, Linear Support Vector Machine, Multinomial Naive Bayes, Logistic Regression. The metrics used for evaluation was accuracy and standard deviation.

Table 5.2 shows the comparisons of model performances over accuracy and the final standard deviations.

Table 5.2: Model Accuracies

Model	Accuracy	Standard Deviation
Linear SVC	77.91	0.005
Logistic Regression	75.81	0.009
Multinomial Naïve Bayes	64.75	0.004
Random Forest Classifier	38.70	0.01

6. Conclusion

The consumer lodge most of their complaints for the financial products they don’t want to buy or use. If a consumer will be contacted repeatedly for products they don’t want, they obviously will be irritated and be forced to put a complaint. The more the complaints data to train the ML models on, the better will be the classification task.

A bigger text corpus will help NLP model to understand the textual context better. e. Natural language Understanding (NLU) is an important step of NLP. The computer must comprehend the meaning of each word. It tries to figure out whether the word is a noun or a verb, whether it’s in the past or present tense, and so on. This is called Part-of-Speech tagging (POS).

A lexicon (a vocabulary) and a set of grammatical rules are also built. The machine should be able to grasp what you said by the conclusion of the process. There are several challenges in accomplishing this when considering problems such as words having several meanings (polysemy) or different words having similar meanings (synonymy), but developers encode rules into their NLU systems and train them to learn to apply the rules correctly.

Thus we can realise that the model created through TF-IDF and Linear SVC classifies most of the complaints into their required arena perfectly.

References

[1] Carsten Giebe, Lennart Hammer Ström, Dirk Zwerenz (2019) : Big Data & Analytics as a sustainable Customer Loyalty Instrument in Banking and Finance
[2] Stefan Nann, Jonas Krauss, Jonas Krauss (2013) : Predictive analytics on public data-The case of stock market.
[3] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, Jochen Papenbrock (2020) :Explainable Machine learning in Credit Risk Management
[4] Durrel Ramratha, Mabutho Sibanda (2014) : Impact of Analytics in Financial Decision Making: Evidence from a Case Study Approach
[5] Corentin Masson, Patrick Paroube (2020) : NLP Analytics in Finance with DoRe: A French 250M Tokens Corpus of Corporate Annual Reports
[6] Sangeeta Mittal, Shivani Tyagi (2019): Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection