**Consumer Complaints Classification**

By- L Abhishek Divecha 1610110193

Rishabh Ranjan 1610110281

**Abstract:**

Consumer complaints are usually informal complaints directly addressed to a company or public service provider, and most consumers manage to resolve problems with products and services in this way, but it sometimes requires persistence.A **consumer complaint** is "an expression of dissatisfaction on a consumer's behalf to a responsible party" (Landon, 1980). It can also be described in a positive sense as a report from a consumer providing documentation about a problem with a product or service.In fact, some modern business consultants urge businesses to view customer complaints as a gift.

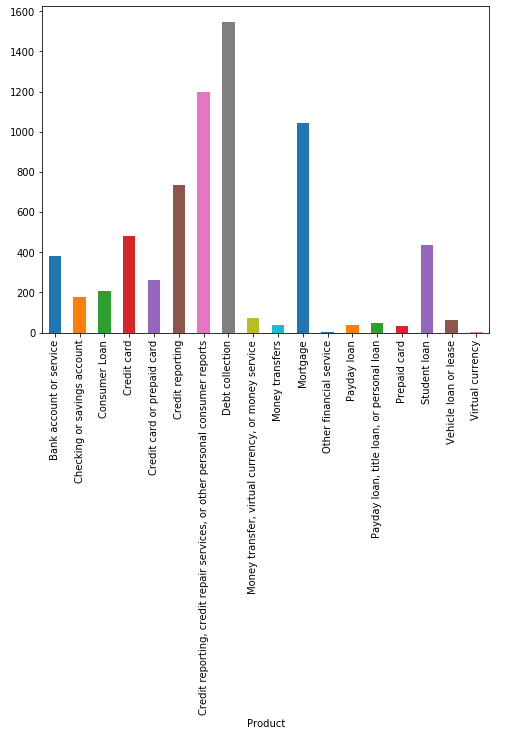
**Keywords:** Consumer, Term-Frequency, Document Frequency, Bag-Of-Words, Linear SVC, Logistic Regression, MultinomialNB, RandomForestClassifier.

**1) Introduction:**

We classify a banks consumer complaints into various categories like : Debt collection, Consumer Loan, Mortgage, Credit card, Credit reporting, Student loan, Bank account or service, Payday loan, Money transfers, Other financial service, Prepaid card etc.

This kind of model will be very useful for a customer service department that wants to classify the complaints they receive from their customers. The classification of the issues they have received into buckets will help the department to provide customized solutions to the customers in each group. This model can also be expanded into a system, that can recommend automatic solutions to future complaints as they come in. In the past, performing these kinds of tasks were done manually by multiple employees and of course, take a long time to accomplish, delaying swift response to the complaints received. Imagine you can classify new complaints with 95% accuracy and route them to the right team to resolve the issue. That will be a win and time saving to any business.

In this project we will classify the consumer complaints into above mentioned categories and train our system with a legit data set.



**2) Related Work:**

Many firms, whether using a paper-based forms system for recording complaints or a more sophisticated software solution, will develop a list or lists of ‘complaint categories’. The complaint category is often developed as a one-dimensional object to combine the products and services offered by the firm with the most common causes of complaint to deliver a single list of options that a complaint handler will be prompted to select from when the details of a new complaint are being recorded (and/or used by quality assurance staff when assessing the types of complaints received during a QA activity).

Sometimes, a firm’s collection of complaint information can be driven by regulatory requirements. Western Australia’s regulator for utility companies provides guidance on the definition and categorisation of water, electricity and gas complaints. The electricity retailers are required by the regulations contained in the guidance to categorise complaints as:

• billing/credit complaints

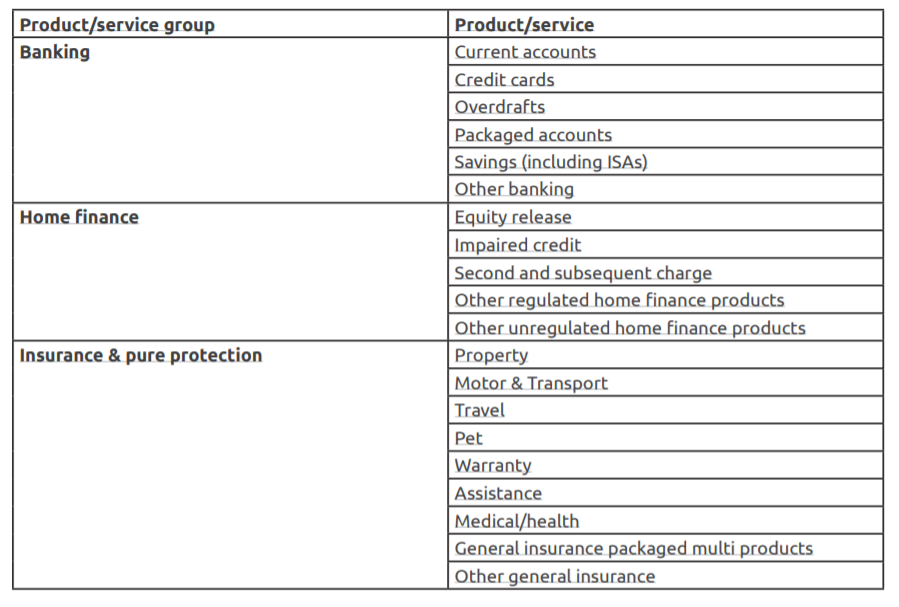
• marketing complaints

• transfer complaints

• other complaints

These are high-level categories. The billing and credit complaints category will of course incorporate a wide-range of lower-level issues: billing mistakes, incorrect billing of fees and charges, failure to receive a rebate, disconnection and reconnection problems and service restrictions due to billing discrepancies. It could be assumed that organisations in this sector will want to classify complaints in more detail than these high-level regulatory categories require.

The current UK regulator, the Financial Conduct Authority (FCA), has requirements that provide for complaints to be categorised based on both a grouping category and a classification of the type of product or service that was the subject of the customer’s complaint. Therefore, the FCA is able - for example - to get an understanding of not only how many consumers have complained about banking products but also how many complained specifically about a current account product.



**Problem vs Root Cause:**

The problem experienced with a product or service differs from the root cause. The root cause may also need to be categorised but this categorisation would need to be captured on closure of a complaint - usually following on from an investigation of a complaint. The categorisation of the root cause might focus upon areas such as ‘Systems’, ‘Procedures’, ‘Human error’ and so on. This differs from the cause that will often need to be categorised upon receipt of a complaint (especially where the complaint is resolved at the point of service within a contact centre or similar environment) but will focus on clear and simple perceptions of the problem experienced (such as delay, staff conduct, failure to do something and so on). Therefore, the root cause category will then provide data that explains how the cause occurred - was the delay due to, for example, human error or systems?

Ideally, you should be aiming to use complaint information to deliver greater insights into the customer experience at each step of the journey that customers take when purchasing or receiving a service or product. ‘Customer journey mapping’ is a very useful method that you can use when categorising and classifying complaints.

**3) Proposed Approach:**

We first start with a analysis of a data set that contains immense amount of data of consumer complaints that is sufficient for training our model, this contains numerous amount of complaints that have been categorised under their respective categories, this will be fed to our software.

**3.1 Data pre-processing and cleaning**

Data requires special preparation before you can start using it for predictive modelling. The objective of pre-processing

is to clean noise those are less relevant to classification such as punctuation, special characters and numbers.

**3.2 Feature Extraction**

**• Bag of Words Model:** We cannot work with text directly when using machine learning algorithms. Instead, we need to convert the text to numbers. We may want to perform classification of documents, so each document is an “input” and a class label is the “output” for our predictive algorithm. A simple and effective model for thinking about text documents in machine learning is called the Bag-of-Words Model, or Bow. The model is simple in that it throws away all of the order information in the words and focuses on the occurrence of words in a document. This can be done by assigning each word a unique number. Then any document we see can be encoded as a fixed-length vector with the length of the vocabulary of known words. The value in each position in the vector could be filled with a count or frequency of each word in the encoded document. This is the bag of words model, where we are only concerned with encoding schemes that represent what words are present or the degree to which they are present in encoded documents without any information about order. There are many ways to extend this simple method, both by better clarifying what a “word” is and in defining

what to encode about each word in the vector.

**• Tf-Idf Vectorizer:** TF IDF is a crude but effective way of distilling documents down to a reasonable and manageable abstraction. However, trade off is this process removes a lot of the semantic information that was in the document.

For unique word in each document, count how many times it shows up in that document. That's “Term Frequency” (TF)

Then, take that unique word and count how many times it shows up in all documents. That's “Document Frequency” (DF)

You can massage DF but next, you divide TF by modified DF (Inverse DF) for each word.

The vectorization part is building a big Vector consisting of each unique word, which is potentially large (tens of thousands of unique words).

Now, if you have a query document, you can build its TF-IDF Vector and find the documents with TF-IDF from the corpus that's “nearest” to the query document. Or use the Vectors to cluster the documents.

**3.3 Model Training**

With this initial data exploration achieved, we are now more familiar with the way data is represented. We are now ready to experiment with different machine learning models, evaluate their accuracy and compare them.

**• Logistic Regression:** A linear classifier, mostly similar to traditional linear regression, but that fits the output of the logistic function. In its vanilla form, logistic regression is used to do binary classification. Multiclass classification with logistic regression can be done through the one vs rest scheme in which for each class a binary classification problem of data belonging or not to that class is done.

**• (Multinomial)Naive Bayes:** A Bayesian model that assumes total independence between features. In our case, this means that P("football") is unrelated to P("stadium"), which of course is a terrible assumption. Still, this model works surprisingly well with the Bag of Words model.

**• Linear SVC:** In this model, we plot each data item as a point in N-dimensional space (where : N is number of features) with the value of each feature being the value of a particular coordinate. Then they are classified by finding a hyperplane that divides the space into two subspaces, one subspace that contains vectors that belong

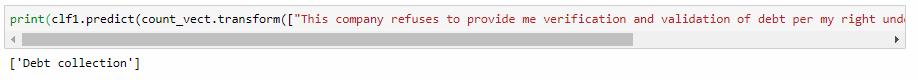
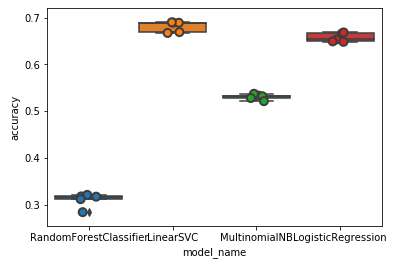
to a class and another subspace that contains vectors that do not belong to that class.

**• Random Forest:** Random Forest (as the name might suggest) is the ensembling of a large number of decision trees, each trained on a random subset of the input features. They work well when complex feature-relations are involved and are relatively robust to overfitting.

**4) Results**

Different classification models resulted in different accuracy measure for the same data set. Naive Bayes model gave a F1 score of 0.53, Logistic regression model gave a F1 score of 0.65, Random Forest model gave a F1 score of 0.31, Linear SVC model gave a F1 score of 0.68. The best performing model was Linear SVC with the highest F1 score.

We use this model for predicting category of complaint.



**5) Conclusion**

A review of Consumer Complaint classification is bestowed in this paper. All the steps i.e. pre-processing, document indexing, feature selection, and Consumer Complaints classification are examined in detail. In addition, stop words filtering using frequency based stop words removal approach is also discussed. In future these algorithms can be tested on larger corpora. Moreover these algorithms can be improved so that efficiency of categorisation could be improved. A combination of algorithm can be used in order to achieve clustering in a faster way.

**6) References**

• Punitha, S. C., and M. Punithavalli. "Performance evaluation of semantic based and ontology based text

document clustering techniques." , Procedia Engineering, 30, 2012,100-106.

• Chen, Chun-Ling, Frank SC Tseng, and Tyne Liang. "An integration of WordNet and fuzzy association rule

mining for multi-label document clustering." , Data Knowledge Engineering, 69(11),2010,1208-1226.

• Hassan, Malik Tahir, et al. "CDIM: Document Clustering by Discrimination Information Maximization." ,

Information Sciences, 316,2015,87-106.

• Li, Yanjun, Soon M. Chung, and John D. Holt. "Text document clustering based on frequent word meaning

sequences." ,Data Knowledge Engineering,64(1),2007,381-404.

• Luo, Congnan, Yanjun Li, and Soon M. Chung. "Text document clustering based on neighbors." ,Data

Knowledge Engineering, 68(11),2009,1271-1288.

• Yan, Yang, Lihui Chen, and William-Chandra Tjhi. "Fuzzy semi-supervised co-clustering for text documents."

,Fuzzy Sets and Systems, 215,2012,74-89.

• Zheng, Hai-Tao, Bo-Yeong Kang, and Hong-Gee Kim. "Exploiting noun phrases and semantic relationships

for text document clustering ,179(13) ,2009,2249-2269.

•<https://www.usefulfeedback.com/wp-content/uploads/2015/09/usefulfeedback-developing-an-effective-guide-FSi-2.pdf>

• Ramasubramanian, C., and R. Ramya. "Effective Pre-Processing Activities in Text Mining using Improved

Porter’s Stemming Algorithm.", International Journal of Advanced Research in Computer and Communication

Engineering 2(12),2013

•<https://medium.com/@opalkabert/classification-of-customer-complaints-using-tensorflow-text-classification-with-word-embeddings-e87952b5147c>

, Electronic Notes in Theoretical Computer Science, 225, 2009,51-65.

• Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128. Springer

7

A PREPRINT - APRIL 24, 2019

• Ali, A. R., Ijaz, M. (2009). Urdu text classification. Paper presented at the Proceedings of the 7th International

Conference on Frontiers of Information Technology