 Automatic Ticket Classification Case Study

# Problem statement

For a financial company, customer complaints carry a lot of importance, as they are often an indicator of the shortcomings in their products and services. If these complaints are resolved efficiently in time, they can bring down customer dissatisfaction to a minimum and retain them with stronger loyalty. This also gives them an idea of how to continuously improve their services to attract more customers.

These customer complaints are unstructured text data; so, traditionally, companies need to allocate the task of evaluating and assigning each ticket to the relevant department to multiple support employees. This becomes tedious as the company grows and has a large customer base.

In this case study, you will be working as an NLP engineer for a financial company that wants to automate its customer support tickets system. As a financial company, the firm has many products and services such as credit cards, banking and mortgages/loans.

Business goal

You need to build a model that is able to classify customer complaints based on the products/services. By doing so, you can segregate these tickets into their relevant categories and, therefore, help in the quick resolution of the issue.

With the help of non-negative matrix factorization (NMF), an approach under topic modelling, you will detect patterns and recurring words present in each ticket. This can be then used to understand the important features for each cluster of categories. By segregating the clusters, you will be able to identify the topics of the customer complaints.

You will be doing topic modelling on the .json data provided by the company. Since this data is not labelled, you need to apply NMF to analyse patterns and classify tickets into the following five clusters based on their products/services:

* Credit card / Prepaid card
* Bank account services
* Theft/Dispute reporting
* Mortgages/loans
* Others

With the help of topic modelling, you will be able to map each ticket onto its respective department/category. You can then use this data to train any supervised model such as logistic regression, decision tree or random forest. Using this trained model, you can classify any new customer complaint support ticket into its relevant department.

Dataset

Let’s first [download](https://drive.google.com/file/d/1Y4Yzh1uTLIBLnJq1_QvoosFx9giiR1_K/view?usp=sharing)the data set.

The data set given to you is in the .json format and contains 78,313 customer complaints with 22 features. You need to convert this to a dataframe in order to process the given complaints.

Starter code

You can download the starter code [here](https://drive.google.com/file/d/1O1AO9AI6Y70e2iqTeD8Ma-hOX6IgQ8V4/view?usp=sharing)

Expected tasks

You need to perform the following eight major tasks to complete the assignment:

1. Data loading
2. Text preprocessing
3. Exploratory data analysis (EDA)
4. Feature extraction
5. Topic modelling
6. Model building using supervised learning
7. Model training and evaluation
8. Model inference

**Note:** Once you have finalised the clusters/categories for customer complaints, the next step is to create a data set that contains the complaints and labels (which you found using NMF). This labelled data set will be used for model building using supervised learning.

You need to try at least any three models from logistic regression, naive Bayes, decision tree and random forest.

You need to select the model that performs the best according to the evaluation metrics.

Q&A

**Evaluation Rubrics**

|  |  |  |
| --- | --- | --- |
| **Stage** | **Weightage** | **Meets Expectations** |
| **Data Reading/Data Understanding** | 5 | The learner has imported the .json data and created the dataframe from it. The learner also has to read the data and make the initial analysis out of it. |
| **Data Cleaning** | 10 | The learner has to do the data cleaning operations like filtering text, removing missing values & renaming column headers. |
| **Data Preprocessing** | 10 | Data pre-processing steps are done using statistical operations like Lemmatization & POS tagging. |
| **Data Visualisation** | 10 | The learner has done the visualisation steps for finding the word count distribution & n-gram distribution. |
| **Feature Extraction** | 5 | Features are extracted from the data using Td-Idf. |
| **Topic Modelling** | 25 | The learner has used Topic Modelling to do the following:   * Set the best number of clusters to 5. * Apply the best number to create word clusters. * Inspect & validate the correction of each cluster for each complaint. * Map the clusters to topics/cluster names. |
| **Model Building** | 20 | Prepared the train & test data using the labels received from NMF. The learner has to apply at least 2 of these models on the dataset created:   * Logistic regression * Decision Tree * Random Forest * Naive Bayes (optional)   Once the models are built the learner has to evaluate their performance using relevant evaluation metrics to select the best model. |
| **Model Inference** | 5 | The learner has applied the best model to predict a  custom text to see its performance. |
| **Code readability and conciseness** | 10 | The learner has done the following:   * Appropriate comments are written wherever applicable. * Overall, the code is readable and has appropriate indentations. |