95-851 / 451 Making Products Count: Data Science for Product Managers

HW3: Natural Language Processing – Fall 2023

Overview

The assignment involves analyzing real consumer financial complaints using NLP techniques, including exploratory data analysis, text data cleaning, sentiment prediction, and leveraging large language models like GPT-3.5 to generate summaries and recommendations. The goal is to gain insights from the CFPB consumer complaints dataset. Do not use ChatGPT or other generative AI tools for steps 1 through 3.2. For steps in step 4 please use OpenAI's API as directed.

Data

The dataset contains consumer complaints submitted to the Consumer Financial Protection Bureau (CFPB) through its website. Each complaint includes fields like the date received, product and issue categorization (e.g. credit reporting, incorrect information), consumer complaint narrative, company public response, company name, state, tags (e.g. servicemember, Older American), and details on how the complaint was handled like timely response and if the consumer disputed. The unstructured text in the complaint narrative provides insights into the specific issues consumers experienced with financial products and services. The metadata fields capture important contextual information like the company, location, and resolution steps. Overall, the rich dataset enables analysis of complaint trends across the financial sector to identify pain points and areas for improvement in customer service.

Tasks

Points in parentheses following the step description indicate the points possible for that step, out of 20 points possible for the assignment.

1. **Tokenize and clean the text** in the consumer complaint narrative field. Profile the data to identify any fields with missing values. Remove punctuation, stopwords, and frequent phrases. Perform stemming or lemmatization to get the roots of the tokens. Identify the most common roots in the cleaned text. (5 points)
2. **Identify the 10 most common consumer complaint topics** based on the product, sub-product, and issue fields Create visualizations to show the most common entries at each level. (5 points)

3. **Sentiment analysis**

Step 3.1: Use Vader to assign a sentiment score for the cleaned text on a scale of 1-5 using the following ranges: (5 points)

1	< -0.5
2	-0.5 to -0.1
3	-0.1 to 0.1
4	0.1 to 0.5
5	> 0.5

Step 3.2 build a model (logistic/linear/random forest/etc.) to predict sentiment on a 1-5 scale using the stemmed/lemmatized words as predictors of the sentiment score. Identify the top stemmed/lemmatized words for each sentiment rating. Display some complaints against each rating and comment on whether they look reasonable. (5 points)

(2 points extra credit) Step 4: Now we shall employ **OpenAI's API and GPT-3.5 model** to ask the following questions:

1. Prompt GPT-3.5 with a sample of complaint narratives and ask it to generate a 1-2 sentence summary of the key issues. Evaluate the quality of the summaries. (.5 point)
2. Select narratives with low sentiment scores. Prompt GPT-3.5 to explain why the customer was unhappy or provide constructive feedback to improve the situation. (.5 point)
3. Prompt GPT-3.5 with the product, sub-product, issue, and sub-issue fields. Ask it to infer the topics and relationships between the categories. (.5 point)
4. Provide GPT-3.5 with the cleaned text for each sentiment rating. Ask it to identify predictive words and themes for each rating. (.5 point)

Submission:

The submission will be a Jupyter Notebook with the name *DSPM_HW3_<your Andrew id>.ipynb*