Technical Report
Group 6

Abhirup, Nipun, Rushikesh, Vamsitej and Tanya

Experiments were conducted in the similar order as mentioned to find the desired results. All the folders and notebooks have the detailed explanation of the code. The following are the different experiments/techniques that were used by the team in order to achieve various results :

1. **Web Scraping**

   Aim: To web scrape reviews and other relevant information from the page.

   Method: We used the help of selenium and Beautiful soup to read the necessary webpage for extraction. We faced great issues in extracting hidden reviews. The later part of the review becomes hidden in case the review provided by the user becomes large. We overcame this issue using the help of chromedriver and its click features.Libraries like selenium, bs4, requests, pandas, time, re and dateparser were used.Date parser was used to parse the dates in readable format.

   Conclusion: We were able to scrape reviews, date, likes,dislikes, ratings and customer name. The output is in the web scraping Final_reviews_excel_total3.xlsx.

   Reference: Web_Scraping.ipnyb

2. **Text Cleaning and Information Extraction.**

   Aim: To clean data

   Method: The reviews are real time data and hence are expected to be untidy. The different forms of text cleaning done on reviews are: Grammar check, Rectify Slang, words, replacing contractions, Spelling correction, Parsing HTML tags etc.

   After the reviews were cleaned, we performed data information extraction on the cleaned reviews. The patterns mined are adverb/pronoun/noun-verb, preposition usage, adjective-noun, adverb-adjective-noun and aspect mining. Also performed duplicated removal between extracts of adjective-noun, adverb-adjective-noun and aspect mining.

   All the necessary methods are mentioned in the folder Text Cleaning and Information Extraction Using Spacy and NLTK.

   Conclusion: Successful data cleaning and spelling correction.

   Reference:Text_cleaning_and_feature_selection.ipynb for adverb/pronoun/noun-verb, preposition usage

Capstone_Aspect_extraction.ipynb for aspect mining

Capstone_Data_Making_adverb_adjective.ipynb for adjective-noun, adverb-adjective-noun pattern extraction

Duplicates_removal_in_adjective_extracted_data.ipynb for duplicate extracts removal

3. **Noun Extraction and Pattern Mining**

Aim: Perform POS Tagging and Extract Nouns

Method: Considering the user features will be in the form of nouns, we split each review into a list of words and performed POS Tagging on them. From the output of POS tagging, we extracted the nouns. To find the pattern and relationships between frequently used, we tried to use Apriori. Apriori algorithm had computational issues when trying to perform pattern mining on 841 extracted nouns. Hence, we went ahead with a more feasible FP growth algorithm

Conclusion: 841 nouns extracted and successful implementation of apriori algorithm.

Reference: Noun Extraction and Pattern Mining.ipynb

4. **Feature Selection**

Aim: To select features for the product.

Method: First step was to clean data and perform some text EDA like stopwords count, avg word count, avg word length in order to understand the data. Then we looked for the words that appeared the most number of times. Following this, we made n-grams =2 and 3 in order to see which words mostly appear together. Later we found out the term frequency score(tf), IDF and TF-IDF. Using the TF-IDF we found out the features.

Conclusion: We were able to find out features like bed, installation, size, delivery, money etc. But later we decided to make features using noun phrases. So for that, we used the apriori algorithm to find out nouns that had the highest support and clubbed these noun phrases under a single feature. Also we perform clubbing of similar looking input phases under single feature phases

Reference: Feature Selection_Group6.ipnyb and Clubbedfeatures.xlsx file in the Polarity and Binary Data Making folder.

5. **Similar Words**

Aim: To find similar words from the data set having the reviews.

Method: We cleaned the data first and removed stop words. Word 2 vec was used to learn vector representation of a target word. We built a model, most similar words where in similar words with the similarity scores was given as output. The main idea was to find similar words for the features extracted so that while processing a review, similar words are also considered.

Another approach was to web scrape similar words from the site synonyms.com. So, we performed both methods and compiled the similar words for each feature together.

Conclusion: We found out similar words for each feature and removed the similar words that did not make much sense or weren't related to the feature in any way.

Reference: Similar Words.ipynb

6. **Polarity and Data Mining**

Aim: Made polarity and binary form datasets, each for phased features and clubbed features

Method: For the phased features, we made a polarity form of the dataset where the column values contain the phased features and the row values contains the polarity values of the information extracted. For the phased feature, finding the feature to which the information belongs was done using cosine method.

For the clubbed features, we did the update using fuzzywuzzy method. The binary form of the dataset updates the row value of each feature into 1 in case the feature is referenced in the review.

Polarity estimation was done using VADER and compound value of the vader was taken.

Conclusion:Polarity and Binary Data making for both clubbed and phased features was done successfully.

Reference: Model_data_making_clean_on_the_clubbed_features.ipynb for clubbed features

Model_data_making_clean_on_the_phased_features.ipynb for phrased features

7. **Calculation of Weights and Clustering**

   Aim: To find weights of each feature and cluster the reviews.

   Method: For calculating weights we used excel. Frequency of each feature was divided by the frequency of the feature in the cluster. K-means clustering was used for clustering the toals of the reviews.

   Conclusion: Found out weights of each feature and then using k means clustering clustered the reviews based on the weights and polarity.

   Reference: weightscalculation.xlsx and weightsfinal.xlsx

8. **Exploratory Data Analysis**

   Aim: To derive meaningful insights from the data.

   Method: Our method was simple i.e to perform simple eda in order to derive insights. The EDA was done on the polarity dataset. We performed eda related to time series as well in order to find out results. Multiple plots were created in order to derive insights.

   Conclusion: Successful EDA helped us in deriving multiple insights about features and their polarities and also insights related to time and pattern. All insights are mentioned in the notebook along with their charts.

   Reference: Capstone EDA.ipnyb and Time_Series_EDA.ipnyb

9. **Front End Deployment**

   Aim: Deploying the review significance results on the front end.

   Method: Flask+HTML along with Heroku was used for the purpose. The code files and all data related to the deployment is in the folder Front End using Flask and HTML.

   Conclusion:The user gets to choose the features, he wants to see the reviews related. He also gets to see the remaining reviews based on significance.

   Reference: main.py file