MDR Report: Trustpilot Sentiment Analysis

Axel Eschholz 100161044

February 23rd, 2024

Approach

**Step 1: Trim Attributes**

The first step is to drop the unnecessary “Date” attribute, as it has no effect on the sentiment expressed in the review.

**Step 2: Adding Heuristics**

To label the data for supervised decision tree training, a lexicon-based sentiment analysis was run on them using the afinn python library. Since the afinn sentiment analysis takes full sentences, this was run before further preprocessing. To add a greater degree of specificity to the results, the afinn sentiment analysis was run on the title and the content for each datapoint.

These scores were then added to the data as “title\_afinn” and “content\_afinn” columns for each datapoint.

A computer screen shot of a computer code

Description automatically generated

**Step 3: Tokenizing**

Using the nlkt library, the text in each review (including the title) is tokenized into individual words. The extracted tokens are then stored in a new column.

A screenshot of a computer program

Description automatically generated

**Step 4: Stopword Removal**

Stopwords, common words with no inherent sentiment, are unnecessary and therefore removed from the token list for each datapoint.

A screen shot of a computer code

Description automatically generated

**Step 5: Vectorization**

From the sklearn toolkit, TfdifVectorizer was used to facilitate Term Frequency – Inverse Document Frequency (TF-IDF) vectorization. TF-IDF is a method used to determine the relative importance of a term in a datapoint in relationship to the rest of the dataset. Applying TF-IDF created new features for each relevant token or bigram. Hyperparameters set the maximum number of produced features, 1000, and the complexity of the n-gram consideration, 2.

A screenshot of a computer code

Description automatically generated

Once the vectorization is complete, the new features are added to the data and the preprocessed data is exported.

MDR I

**Initial Dataset**

The data consists of 2000 customer reviews in plaintext. The text contains bad spelling, symbols, and messy punctuation. The heuristic value for the title and content is calculated using the afinn sentiment analyzer.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Description | Usage |
| index | continuous | index of consumer review | ID |
| Title | text | Title of the customer review | input |
| Contents | text | Contents of the customer review | input |
| Date\* | temporal | Date review was created/updated | N/A |
| title\_afinn | interval | Heuristic analysis of title | output |
| content\_afinn | interval | Heuristic analysis of content | output |

\*Dropped due to irrelevance within this project

MDR II

**Preprocessed Dataset**

Over 1000 input features across the same 2000 data points based on the TF-IDF analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Description | Usage |
| index | continuous | index of consumer review | ID |
| [TF-IDF] | continuous | TF-IDF weight of each token | input |
| Title | text | Title of the customer review | input |
| Contents | text | Contents of the customer review | input |
| Tokens | nominal | Tokens extracted this review | input |
| title\_afinn | interval | Heuristic analysis of title | output |
| content\_afinn | interval | Heuristic analysis of content | output |