|  |  |  |  |
| --- | --- | --- | --- |
|  | Definition of functions | input | output |
| Text Preprocessing | Combining NLP steps | Textual content with label | The bag of word matrix |
| ML preprocessing | Imbalance data textual content |  |  |
| Training | Machine learning training | Training machine learning algorithms | The trained models (confusion matrix) |
| Comparison | Comparison and visualization of algorithms | Confusion matrix | Visualization and comparison |
|  |  |  |  |

Works steps

|  |  |  |
| --- | --- | --- |
| Joe |  |  |
| Zara | Draft functions |  |
| Atousa |  |  |

**To-do-list:**

Read through already existed packages and do comparison with what we are planning to do

GitHub

Blog

Writing a documentation of each function

Write the functions

Checking

Testing

Package

Shinny app

Final documentation

Summary/abstract

Sentiment analysis of consumer reviews

Consumer reviews are very instrumental in projecting the development and direction of businesses. The wealth of information provided with online reviews can be analyzed to access the general sentiment of the consumer towards the product/service. This sentiment analysis will provide general idea and quantitative measure of the consumer’s sentiment. Consumer review with NLP of the text review will generate a lot of important insight that a 1-5 scale don’t. It is a way to get a sense of the consumer review from the wealth amount of data that online consumer review provide nowadays.

The required data set, in .tsv format, consist of a text review and a binary input of 0 or 1 indicating whether the consumer liked the item or not. These will be balanced using “ROSE” package, which attempts to generate synthetic data by randomly over-sampling examples in case of imbalance data.

Then using a text mining package “tm”, it will mine for the words/word base of interest such as keywords that conveys sentiment based on dictionary list provided by the package. This “stream-lined“ tokens that were independent variables on the raw dataset is then joined in a matrix through either of the three algorithms of the user’s choice such as Bag of Words, Tfdfi, or Bigram and will be analyzed based on its sentiment. Next, using “caTools” package, we split the data set to 0.2 data and 0.8 training set split ratio to be fed to different classification models of provided by the existing machine learning algorithms training models.

The package also compares the accuracy of the prediction model based on the machine learning algorithm which the user can choose from: decision tree, naïve bayes, or random forest. The result is different parameters of the sentiment prediction that is built in a confusion matrix: Accuracy, Precision, Recall, and F1 score. Lastly, these parameters will be visualized as a wordcloud as well as presented in an interactive Shiny app. Overall, this package will put together different algorithms and visualization of a version of natural language processing (NLP) and will allow a broader audience of user to access this tool.