Machine Learning Engineer Nanodegree

Capstone Proposal

Prabhakar Mudliyar
July 20th, 2019

Proposal

Domain Background

In this current market to improve our product and increase the product purchase, the main goal now to many companies is to study the customer. The main focus is now on how customer accepts the product and provides the feedback. This feedback will only help the product to improve. In order for this we need to study the customer feedback, which comes from a lot of different places and people which cannot be done manually. Hence we make use of Supervised Learning Algorithms which help understand the customer sentiments while they review the product.

Why use Supervised Algorithms, as this will help to automate the process of reviewing the large datasets manually and also reduce human errors. And also in time how much we train our algorithm with data it will only increase performance.

Kaggle Dataset: www.kaggle.com/sid321axn/amazon-alexa-reviews

Problem Statement

Amazon has different type of products Alexa products and they want to know the feedback from the customers for the products. The objective is to perform the analysis of customer sentiments and derive insights into consumer reviews.

The Kaggle data consist of more than 3000 customer reviews, star ratings, review date, variant and feedback of Amazon Alexa products. The objective is to discover insights into consumer reviews and perform sentiment analysis on the data.

Datasets and Inputs

Kaggle Dataset: www.kaggle.com/sid321axn/amazon-alexa-reviews

The above dataset contains amazon_alexa.tsv which is required for predicting the sentiment analysis. As this is a Kaggle Dataset the data was available on the website. The dataset consists of columns mentioned below

- Rating: Number ranging from 1 to 5 indicates likeness of the product by the customer.
- Date: On which date the review was recorded.
- Variation: One of the variants of Alexa Products.
- Verified_reviews: The comment or the review given by the customer for the product.
- Feedback: This is the target variable which help to find the accuracy

Solution Statement

The goal is to find the sentiment from the review provided by the customer. The main problem being the review which consist of lot of vocabulary and words which make it very difficult to predict the sentiment. So to make this easier will try to use of one of the Supervised Algorithm called Random Forest Classifier (Decision Trees). To solve the problem of vocabulary we make use of Feature Extraction method CountTokenizer which covert a collection of all words to a matrix of token counts. The matrix will contain all the unique words from the review as column and it will be defined by the count of how many times the word occurs in the document. We also the feature encode the variant column to remove dummies.

Benchmark Model

The benchmark model will be the one which provide the maximum accuracy. For this we will try and check the accuracy with Confusion Matrix to find the benchmark. To improve the model we can add more rows to the dataset like the length of the review which may improve the performance.

Evaluation Metrics

To evaluate the algorithm, we make use of the Confusion Matrix which provide the accuracy. The classification report will also show the results of the precision, recall, f1-score. Based on these scores we will be evaluate our model better.

Project Design

To predict the sentiments analysis of the Amazon Alexa products, first will start the with loading and visualizing of different feature to understand our data. After that, we can remove the features which will not helpful in the algorithm to reduce the overhead.

Based on the features, first will implement this with Decision Trees (Random Forest Classifier.

Decision Trees are supervised algorithms where data is split according to a certain condition. DT consist of nodes and leaves where leaves are the decisions or the final outcome whereas the decision nodes is where the data is split based on a certain attribute.

Random Forest Classifier is a type of ensemble algorithms, it creates a set of DT tress from randomly selected subsets of training data and then it combines the votes from the different DT tree to decide final class of test object.

To reduce the overfitting problem, we will use the Random Forest Classifier.

Now we will use the feature extraction methods to convert the vocabulary in the review feature to more logical understandable numbers which will help in prediction.

Once this is done, we can split the data into train and test datasets. After applying all the steps, we will apply the confusion matrix to evaluate the accuracy.