|  |
| --- |
| Owner: Asim Iqbal  Email: asimiqbalsiddiqui.1@gmail.com |
|  |

|  |
| --- |
| **Artificial Intelligence | Sentiment Analyzer** |

****

|  |
| --- |
| SENTIMENT ANALYZER |
| Project Overview: Sentiment analysis is a common natural language processing (NLP) task in the field of Data Science. It is the process of determining a context of a piece of writing whether it is positive, negative or neutral. In this project, I will train the model with such data and algorithm to make it able to predict the sentiment of a customer review on movies with good accuracy. Standards:  * Data Science Models. * Data Science Algorithm. * Natural Language Processing.  Project requirements: For this project, following things are required:   * Testing Data (positive review + negative reviews). * Training Data (positive review + negative reviews). * Natural Language Processing Model. * Python Libraries.  Scope: The project will only focus on movie reviews and not on others like product reviews. Further it will only focus on processing text in English and not in any other language. WORKING:SETTING UP RESEARCH GOAL: The goal of this project is to predict the viewer’s sentiment for the review he/she gives for the movie. This project is important for the movie producers to analyze the viewer’s views for the movie which help them understand content requirement deeply. RETRIEVING DATA: The data used to train and test the model is compiled by [Andrew Maas](http://ai.stanford.edu/~amaas/) and is collected from [IMDb Reviews](http://ai.stanford.edu/~amaas/data/sentiment/).  The data contains 50,000 movie reviews. The first 25,000 are separated for training purpose and the rest 25,000 are for testing purpose. In each 25,000 collection, 12,500 contains positive movie reviews and 12,500 for negative movie reviews. PREPARING DATA: Before processing and taking data to production environment, it is mandatory to process the data first. Below are the methods used to process the data: **3.1- DATA CLEANING:** The raw text is messy for these reviews. So before performing any analytics, data cleaning is required for the collected data set.  To clean the data, REGEX EXPRESSIONS are used. Regex is the general formula that will run over data set to perform desired set of output. The characters such as comma, brackets, inverted commas etc. are the unnecessary for any movie review. Moreover, some other characters such as line breaker, unusual white spaces, slashes etc. are also useless. Following this use case, regex expression is written to clean the data set. **3.2- DATA TRANSFORMATION** Data transformation is the process to transform data from one state to another. Since the model that is being used in this project is Logistic Regression (will be explained later in this report), which works on numerical data set. But, the data we have is currently simple English.  So, for this data to make sense to our model algorithm, we need to convert the data into numerical form. For this purpose, vectorization is used.  Vectorization in data science is the process which is used to map character data into numerical form. This method creates a very large matrix, which contains the count of every unique word in the data (data in our case is movie reviews). **3.3- COMBINING DATA:** After transforming data into numerical form, we will combine the data to process further. Python ‘.zip()’ method is used to combine the dataset. This method will map the data in a sense that each word will correspond to the column consists of its number of occurrences in whole data set. Model: Model is used to train the machine and make it able to predict the new data based on the training. In this project, Logistic Regression Model is used. Logistic Regression is a good baseline model in our case because it’s easy to interpret, tends to perform good accurate results compared to other models on such sparse data and is also tends to learn quickly.  Logistic Regression works to find the squiggle shape of the data set. Unlike Linear Regression, Logistic Regression does not find a straight forward regression line to predict the outcomes but works on to categorize the data in a binary manner of 0s and 1s or Yes(s) and No(s).    After categorizing the data, the model will try to fit the squiggle curve over data as good as possible using different values of Hyper Parameter(C). In this project, hyper parameter values that are used to find the best possible accuracy are 0.01, 0.05, 0.25, 0.5 and 1.0. The value 0.05 is found to be fitting accurately compared to other values. Following are the obtained results:   |  |  | | --- | --- | | HYPER PARAMETER VALUE | ACCURACY | | 0.01 | 0.874 | | 0.05 | 0.883 | | 0.25 | 0.880 | | 0.5 | 0.878 | | 1.0 | 0.876 |    TRAINING: Now that we have selected the model, we will train the model using the data we collected and cleaned earlier. With the optimal hyper parameter value, we will use 25,000 reviews for training. Since the data is based on numerical values only and not any complex data, it won’t take much time for Logistic Regression Model to train completely. TESTING: After training out model completely, we will test our model with the remaining 25,000 data set. It is necessary to test the data to ensure that the model is not in over fitting or under fitting condition.  The results of testing show good accuracy of model. The Model reached the accuracy of up to 91% overall while predicting. Below are some tested results for positive and negative reviewed words:   |  |  | | --- | --- | | POSITIVE WORDS | PREDICTED ACCURACY | | Excellent | 91.2% | | Perfect | 79% | | Great | 65.7% | | Amazing | 61.6% | | Superb | 60% |  |  |  | | --- | --- | | NEGATIVE WORDS | PREDICTED WORDS | | Worst | 90.2% | | Waste | 87.1% | | Awful | 82.6% | | Poorly | 73% | | Boring | 65% |  Conclusion: This project has successfully predicted the sentiments for the movie reviews. The model has predicted outcomes with good accuracy. NLP processing is an important subject in the growing world. With the help of this model, we can now help production companies analyze the viewer review more quickly and accurately for better production of movies. |
|  |