Ecommerce with amazon reviews (Kindle dataset)

Our project Amazon reviews on kindle dataset use customer feedback reviews or ratings on the product, where we utilize techniques such as opinion mining, text mining and sentiments, which has affected the surrounded world by changing their opinion on a specific product. Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information from the user/customer, such as opinion and attitudes, from language, helping in finding the mood of the customers about a purchasing of a particular product or topic in relation to entities such as products, services, organizations, events, topics and their different attributes. The whole process includes summarization in three steps which are as follows:

(1) Product feature-based, which is given by the customer.

(2) Identify expected features in each opinion sentence or review

(3) Finding out whether the feature/opinion is positive, negative, or neutral and finally, the summary will be created.

(a) assumptions made if any

We went by applying stemmed algorithm because lower features mean more information density in the compressed columns. Because it was found that the stemmed algo has features that were 18% lower than that of lemmatized approach.

(b) the approach used -a summary

Consumers want to find useful reviews as quickly as possible by the help and assistance of the rating system. Therefore, models that are capable to predict the user ratings from the text review are critically important. Getting an overall sense of a textual review could in turn improve the consumer experience. Also, it can help businesses to increase sales, and improve the product by understanding customer’s needs. The dataset was pre-processed to reduce the missing and the null values and different visualizations carried out to interpret the data for fruitful results. NPS (Net Promoter score) was calculated based on the ratings of the user and the graph turned out to be the one as expected.

Finally, the model was built to find the accuracy and recall on the dataset to carry out sentiment analysis and draw necessary interpretation.

(c) summary of the results reported

The model turned to give higher accuracy and precision of 98% each with a recall of 99%.

(d) any limitations reported?

We tried to check if all ratings were given by the same user or were, they any discrepancy due to the same. This was sorted after visualizing and comparing with the help of graphs like bar graph and pie chart.

(e) any lacuna in their approach/ evaluation that you inferred?

We also carried out time-series analysis on reviews and check it for small to large intervals of time.

Proposed problem statement with the specific issue you intend to address

Our aim in the project was to draw sentiment analysis by taking feedback from the customers on amazon kindle reviews dataset and infer necessary information to extract the requirements of the user to best fit their needs. In today’s world with increasing advance technology and never-ending demand of the users, it becomes extremely essential to built and predict essential models that can help curb the requirements of the user and accommodate with the best-suited services. We built our model to increase the efficiency and accuracy to increase the returns for the business and smoother experience.

How is your approach (or the type of problem you are looking at) different from what has already been done? (or if you are attempting to improve upon someone else's work, explain in what way it distinguishes itself from what has been reported)

By drawing and visualizing insights from the dataset on various parameters and compare the results obtained earlier and improve it was the focus we tried to achieve in the project. Better plots, graph and chart helped to adopt strategies to increase the performance of the model and earn better results on future data also. We also incorporated methods/functions like lemmatization and stemmed process, calculate the NPS score and carried out an analysis on time series data to check and validate the dependency of our dataset in accordance with time.