**AMAZON REVIEW DATASET**

Project Documentation



**Submitted by –**

Nitesh Singh

Aditya Kumar

Varun Verma

Anurag Pawar

**Project Introduction -**

The year was 1994 when Bezos launched Amazon out of his garage. In 1995, the first product was launched by Amazon. It was a book that was sold to 50 states in 45 different countries within 30 days (Oberlo 2021).

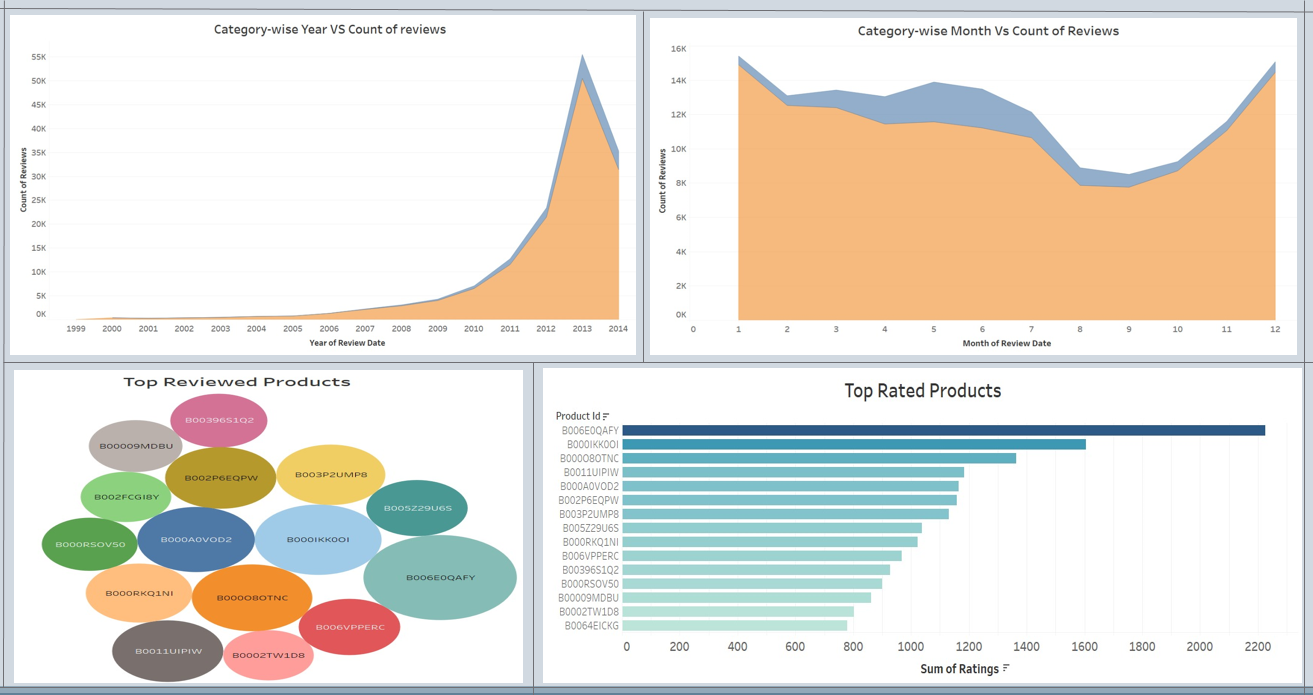
Within 26 years, Amazon holds the title of the world’s largest online retailer and has become a household name. Amazon has become synonymous with online shopping and continues to grow by developing new products, acquisitions, and different service offerings to enlarge the customer base

**PRODUCT CATEGORIES –**

Tools And Home Improvement,

Patio, Lawn and Garden

**Exploratory Data Analysis –**

****

1. By using the count of reviews and Year column in first graph we can interpret that there is an increasing trend in reviews after 2011 – 2013.
2. By using the count of reviews and month column in second graph we can interpret that most of the reviews are given between march till June.
3. B006EOQAFY product is top rated and top reviewed product and has been plotted using bubble chart and Horizontal bar graph.

**Data Pre-Processing -**

1. Column names were changed for the better understanding of the viewer and column named unixReviewTime was converted into date time format by using to\_datetime function and a separate column was created.
2. By using this column month and year of the reviews were noted into a new column.
3. By using nltk library stopwords were removed from the dataset.
4. By using nltk library all special characters such as <. \* ? > were removed.
5. By using nltk library lammetization was carried out on the dataset.

By using above mentioned methods Data Pre-Processing were carried out.

**Natural Language Processing (NLP)**

1. By using Textblob sentiment analysis was carried out and polarity was found for the reviews of products and they were segregated as positive, negative and neutral.
2. By using the polarity and score column in a for loop various negatively rated products were found with there Product id such as B003VPMJGE (Cutting Blade) or B0052W5OP2 (Animal kill Traps).
3. By using same method and review text column various negative reviews were analyzed and few reasons were Poor quality product, Leakage in product, Poor built, Items missing etc.

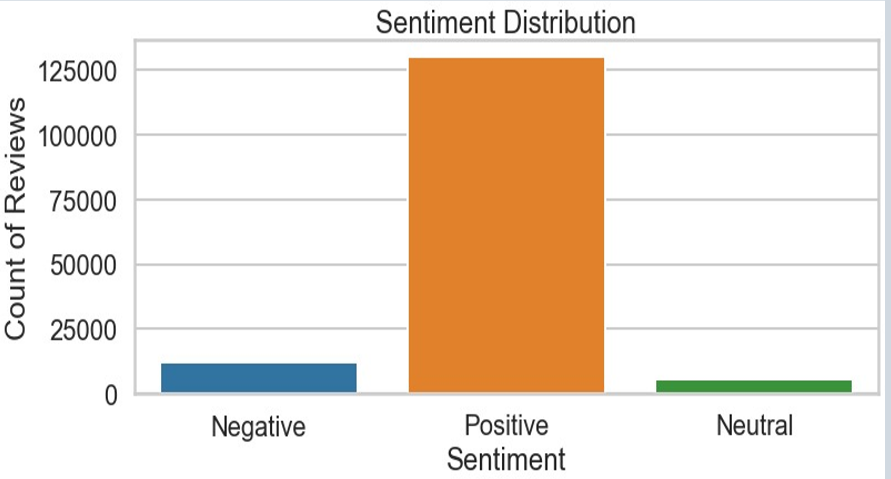
By using above methods NLP and sentiment analysis were carried out on dataset.

**Model Evaluation -**

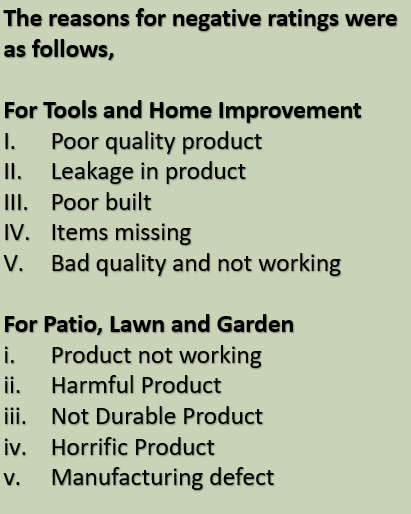
1. In forecasting firstly, we done resample with sum.
2. Take the necessary column and convert Sentiment column into Label Encoder.
3. Applying Decomposition method to check trend and seasonality in given data set.
4. Then need to see if there is trend and seasonality is present in dataset given. Need to make stationery i.e., no seasonality and trend should be present there at time of prediction. Mean a covariance would be constant.
5. Select the order of (p, d, q) for the model that is giving min. AIC and LL.
6. Making data stationery need to used model ARMA, ARIMA, SARIMA. to predict the data and compare with actual data.
7. Calculated the RMSE value of all the models to check which model is good for forecasting.
8. In last, forecast the data on ARIMA model that is giving positive result and minimum RMSE.

**Problem Statement –**

1. Based on the ‘review text’ and ‘summary’ of the data, identify how a review can be classified into a particular category (for example, positive, negative and neutral). Analyze the text well to build the categories. In case of many negative sentiments associated with a product, try to find out the reasons.



Above graph shows the representation of sentiments.

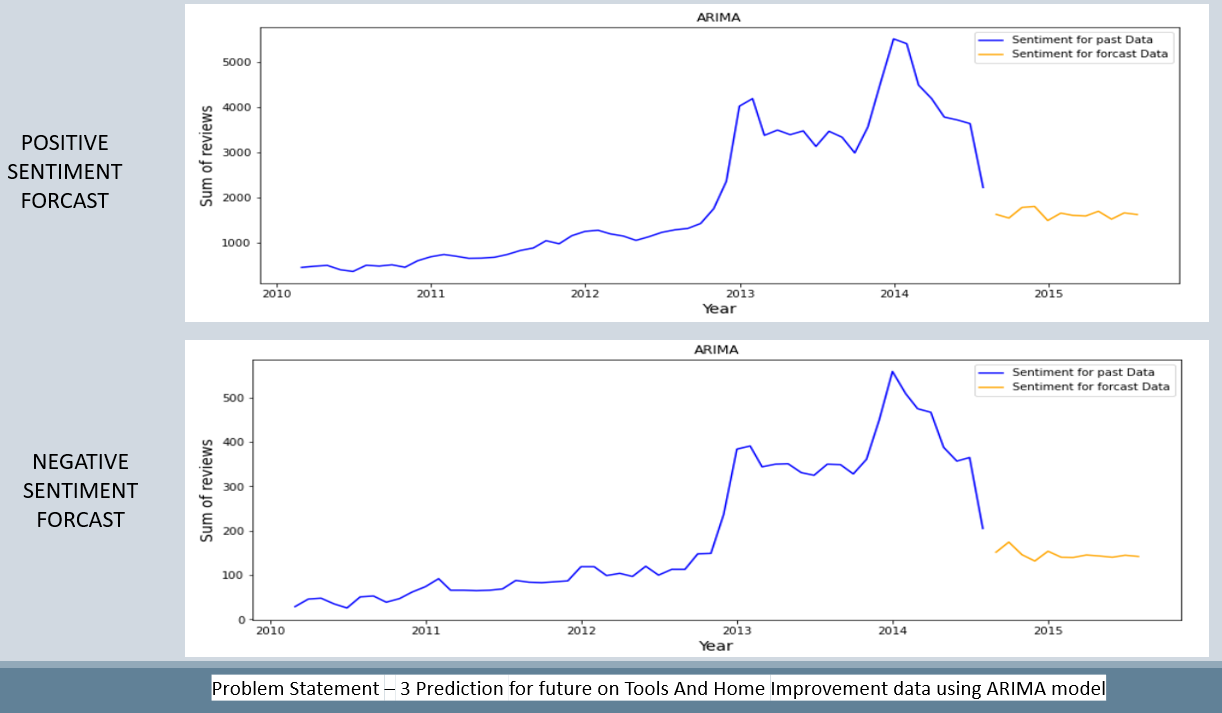


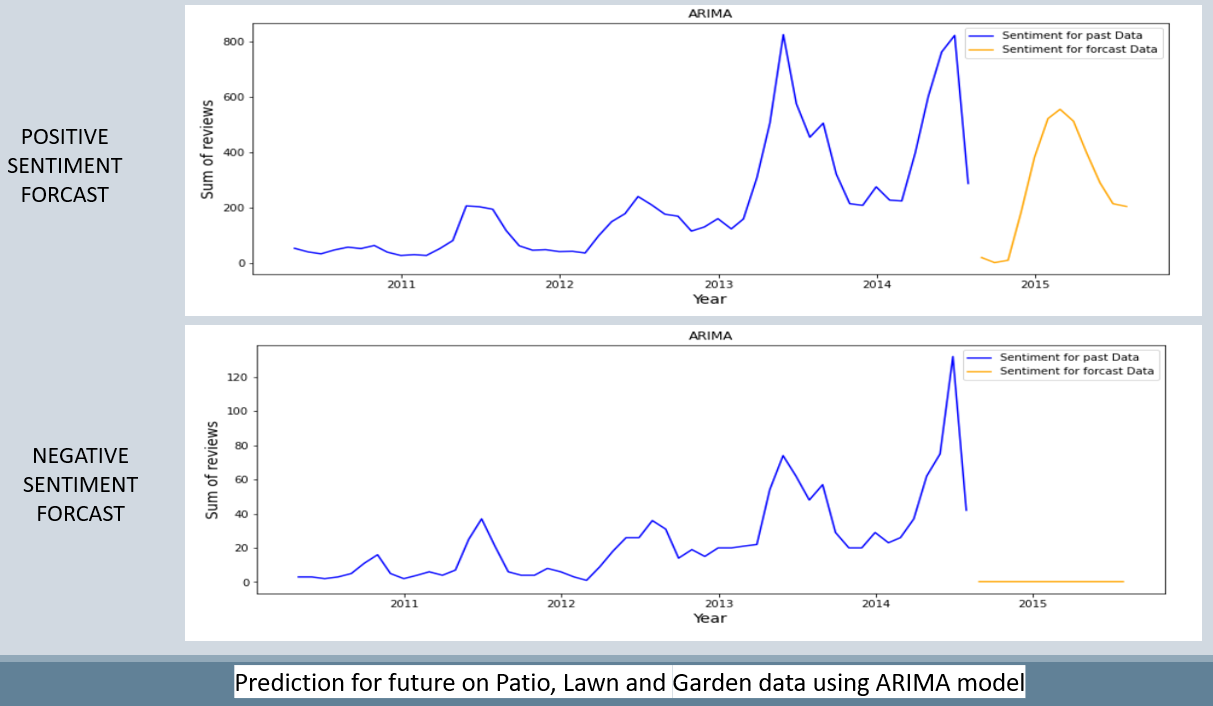
Above picture shows the reasons for negatively rated products.

1. Identify the names of few products by analyzing your input text data.



1. Predict future data trends: How the sentiments of reviewers change with time. Suppose input data is provided for the period 1996-2014. Your task is to predict the trends after 2014.

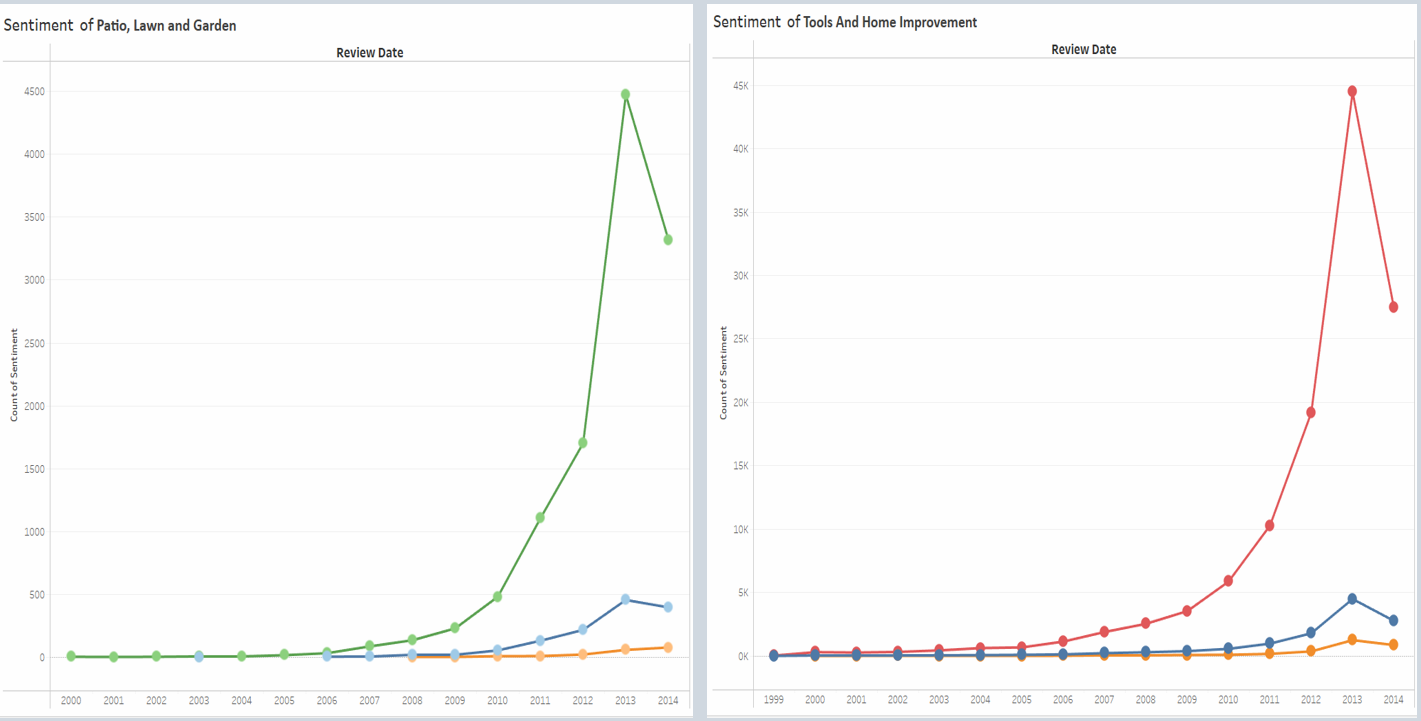




1. Can you find any relations/buying trends of customers/any other interesting analysis between given pair of categories (Example, Beauty and Clothing categories).

During our analysis we didn’t found any customer or product which were in both the data sets but we realized that the problem faced by the customer were somewhat similar.

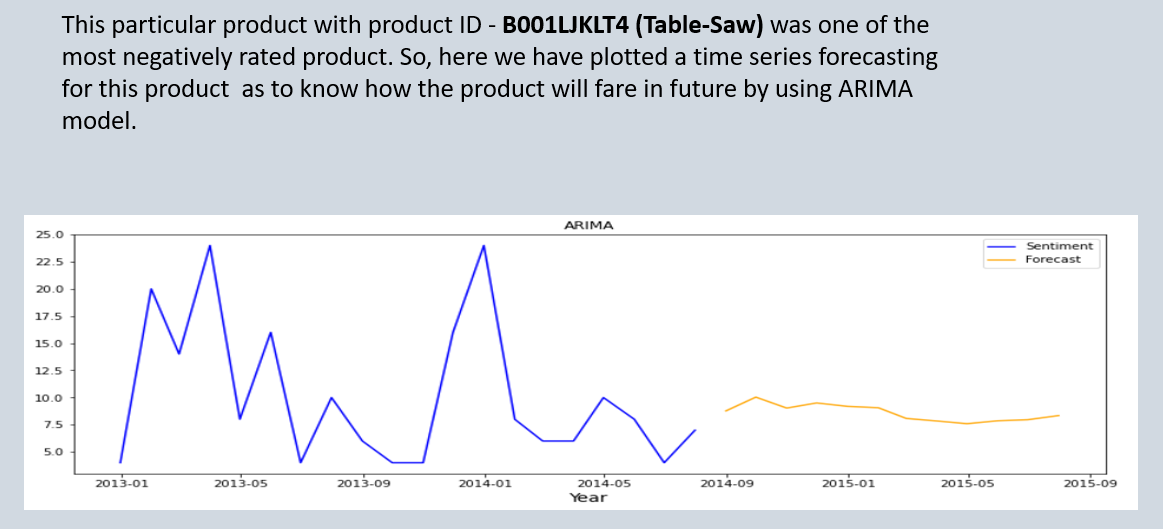
Regarding the trends in Tools and Home improvement dataset the reviews were at peak in 2014, but afterword's it Declined considerably, whereas Patio, Lawn and Garden dataset reviews were at peak even after 2014 but it had also a steep fall.



**Conclusion –**

* The sentiment analysis shows that most of the reviews are positive for all products followed by negative reviews and the neutral reviews being least in number.
* In ARIMA model the value of actual v/s predicted are almost similar than other models.
* The Year wise pattern of sentiments trend is similar for both categories.

The following graph shows the prediction of a particular (Table-Saw) negatively rated product.

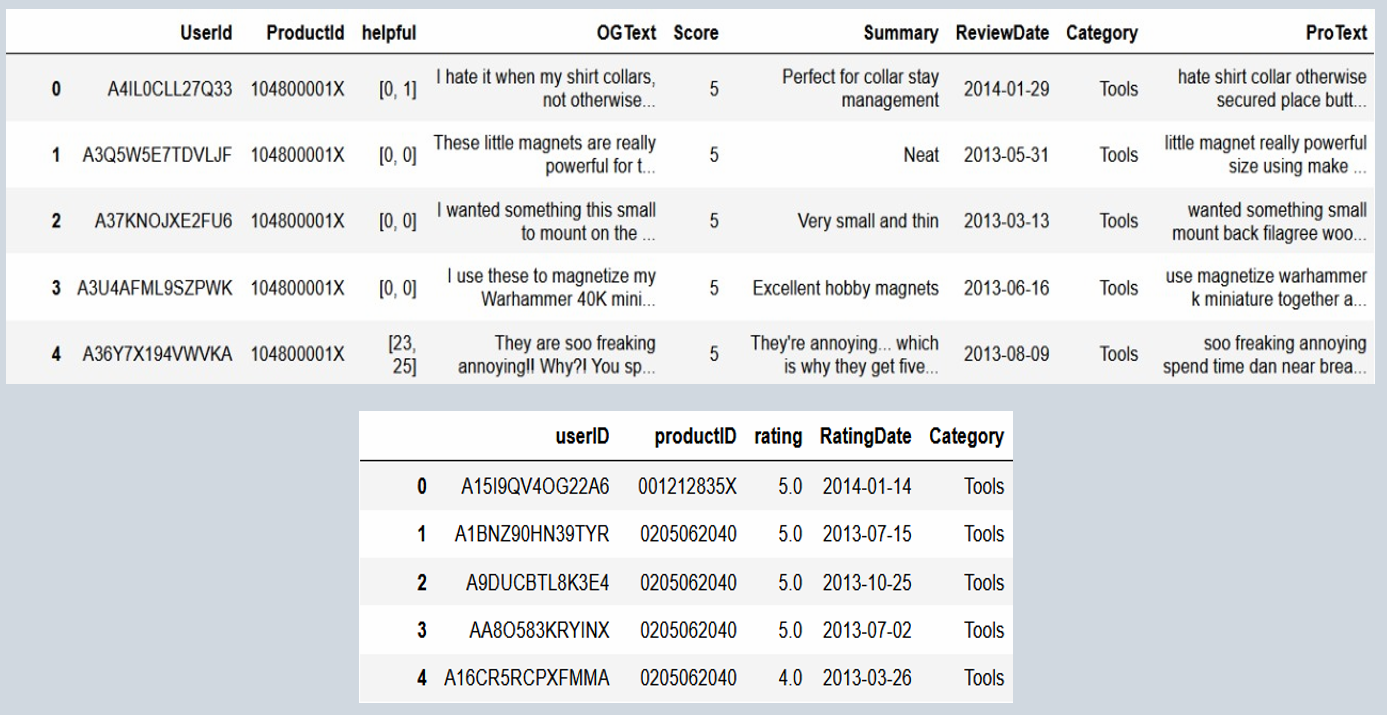


**Prescriptive Analysis –**

* Few of the reasons for those negative sentiment were : Over Packaging, Poor Quality Products, Items Missing or Manufacturing Defects.
* So, we recommend in improvement of shipment and delivery services, to make the warehouses more robust and crosschecking to verify the missing items in the delivery.

**Appendix –**

Test output data –



Dataset –

* Tools And Home Improvement (rating & reviews)
* Patio, Lawn and Garden (rating & reviews)

**References –**

* <https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775>
* <https://en.wikipedia.org/wiki/Time_series>
* <https://towardsai.net/p/nlp/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0>
* <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>