Sentiment Analysis of Amazon Reviews

KIRTI CHAUDHARI

Goals

Analyze	Problem Statement: Analyze Amazon review dataset for Electronics category to find out sentiment analysis.
Visualize	Understand the overall trends and visualize the dataset.
Apply	Apply Machine learning on the dataset to predict sentiment analysis based on the review text

Data Acquisition

- Source : https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz
- Electronics Category
- 130+ million customer reviews since 1995

Column Name	Description
marketplace	2 letter country code of the marketplace where the review was written.
customer_id	Random identifier that can be used to aggregate reviews written by a single author.
review_id	The unique ID of the review.
product_id	The unique Product ID the review pertains to. In the multilingual dataset the reviews
product_parent	Random identifier that can be used to aggregate reviews for the same product.for the same product in different countries can be grouped by the same product_id.
product_title	Title of the product.
product_category	Broad product category that can be used to group reviews. (also used to group the dataset into coherent parts).
star_rating	The 1-5 star rating of the review.
helpful_votes	Number of helpful votes.
total_votes	Number of total votes the review received.
vine	Review was written as part of the Vine program.
verified_purchase	The review is on a verified purchase.
review_headline	The title of the review.
review_body	The review text.
review_date	The date the review was written.

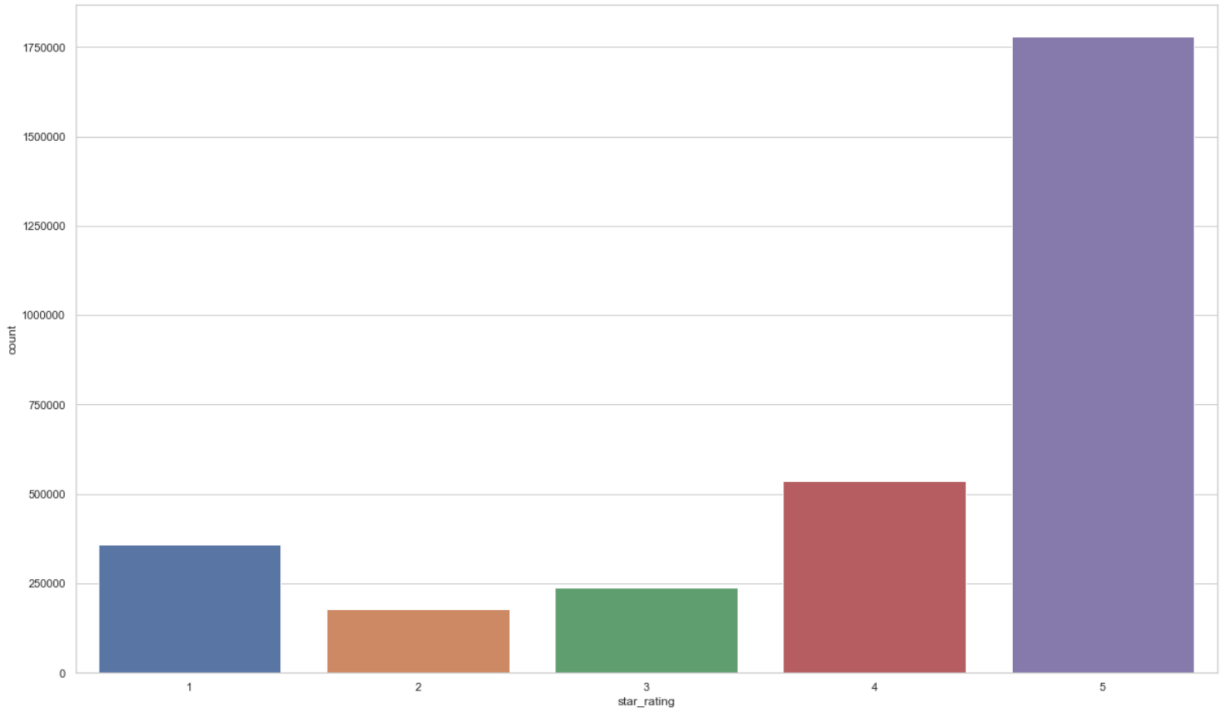
Data Preparation and Cleaning

- Check empty/null values
- Drop Null records
- Correct Datatypes for columns
  - Review_date– string to datetime

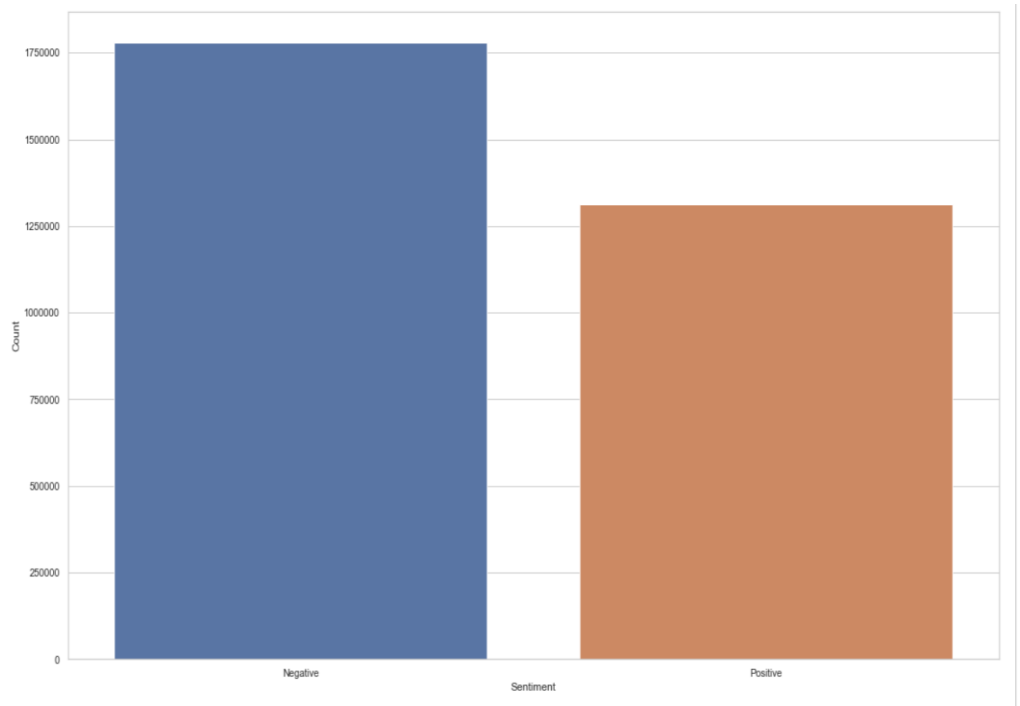
Data Wrangling

- Create New column – Sentiment
- Star_ratings >= 4 -> Positive , else Negative

Data Visualization

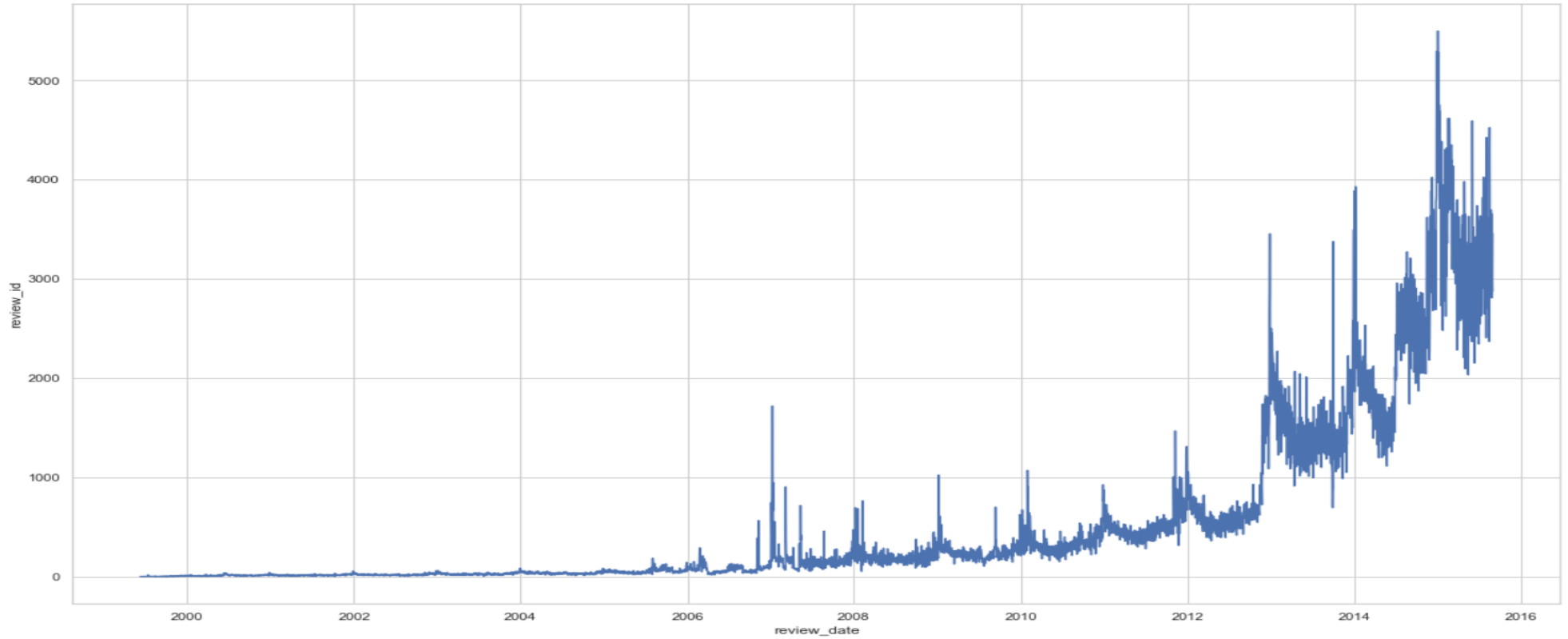


Review Count for Each Star Rating



Sentiment distribution of Dataset

Data Visualization

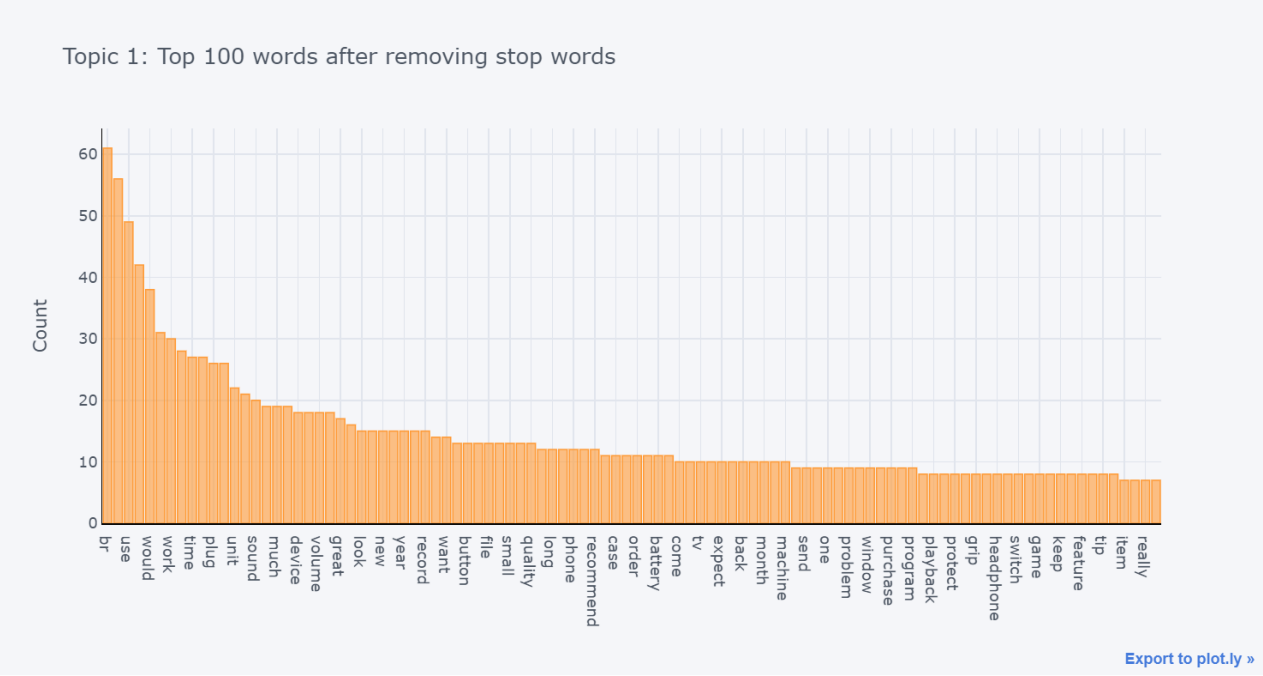


Monthly Trend of Reviews

Data Modelling

- Primary Goal : NLP analysis and Binary classification of Review body
- Use Pycaret library to do NLP Sentiment Analysis
  - Setup the environment
  - Load the data
  - NLP - Create, Assign , Plot, Evaluate Model
  - Classification – Compare, Create, Tune, Plot, Evaluate Model

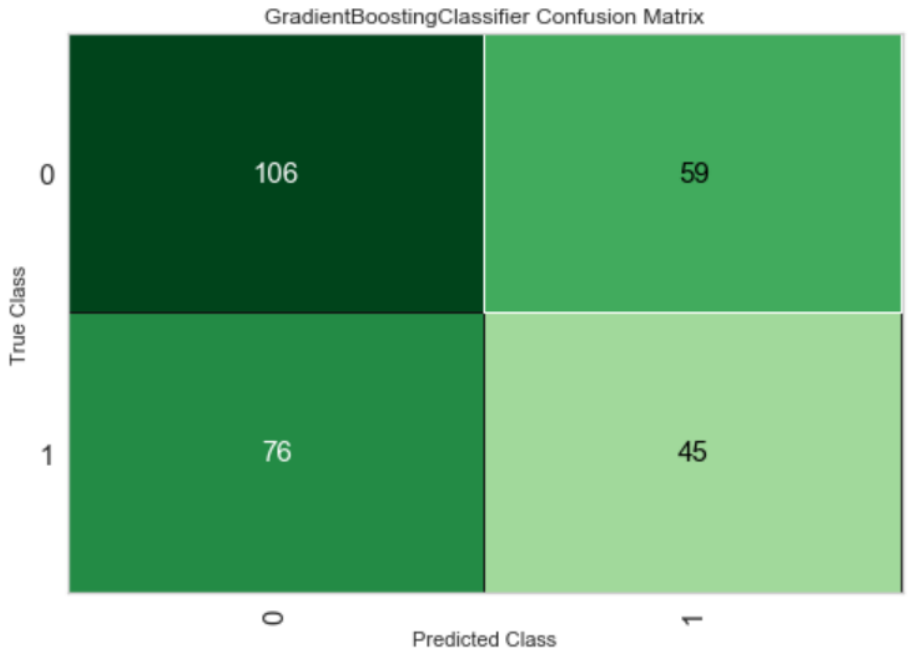
NLP Analysis



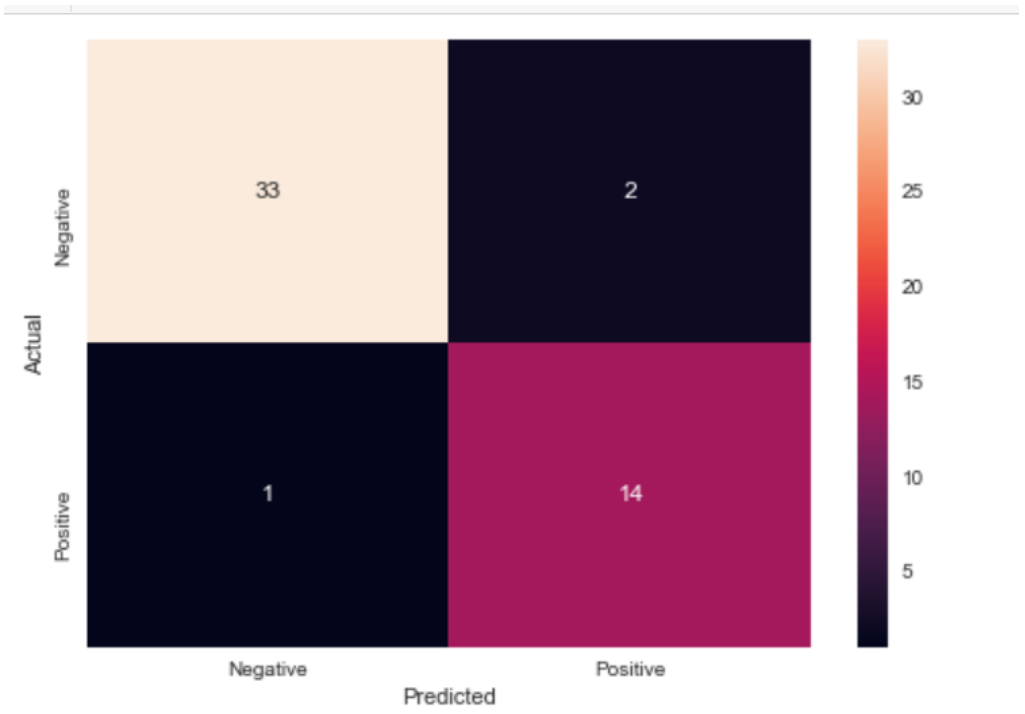
Top 100 words after removing stop words

Word Cloud for Topic 2

Binary Classification



GBC Model – Confusion Matrix



TF-IDF Logistic Regression

Conclusion

- Data acquired Amazon
- Data cleaning, wrangling steps were performed on the dataset
- Used PyCaret NLP analysis on the dataset.
- Used PyCaret Binary Classification.
- Also used Sci-kit learn TF-IDF Logistic Regression.
- Model provided 69% Accuracy rate.

References

- Project Github:

https://github.com/kirti-chaudhari/SpringBoard_DataScience_Career

- Jupyter Notebook :

https://github.com/kirti-chaudhari/SpringBoard_DataScience_Career/blob/master/CapstoneProject2/Capstone%202%20-%20NLP.ipynb