Springboard Capston-2

Topic: Mining Amazon Product Reviews using NLP

Reeshma K

Problem Statement

- Ratings alone do not give a complete picture of the products we wish to purchase. So secondary option is looking at the reviews. Review’s plays an important role in the decision-making process. If the number of reviews is less, it is easy to read and understand but what if there are thousands of reviews.
- So, the problem is how we can analyze great number of online reviews using Natural Language Processing (NLP).

Purposes

- Help Consumers to understand sentiment of the review (Sentiment Analysis).

- Help the consumers to get consumer feedback in the form of topics covered by the reviews without having to go through all of them(Topic Modeling).

- Enable consumers to quickly extract the summary of the reviews without reading the entirely(Text summarization).

Data Description

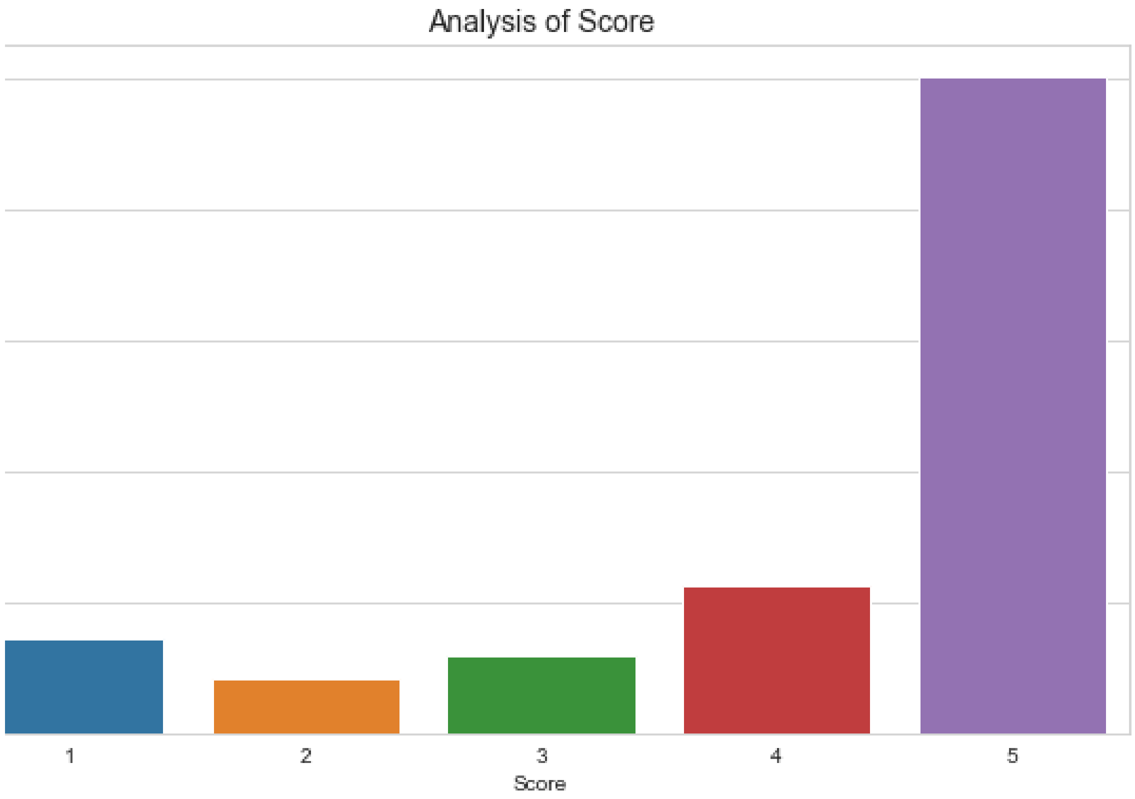
- **Dataset:**https://www.kaggle.com/snap/amazon-fine-food-reviews

- **Context:** This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

Text Wrangling and Pre-processing

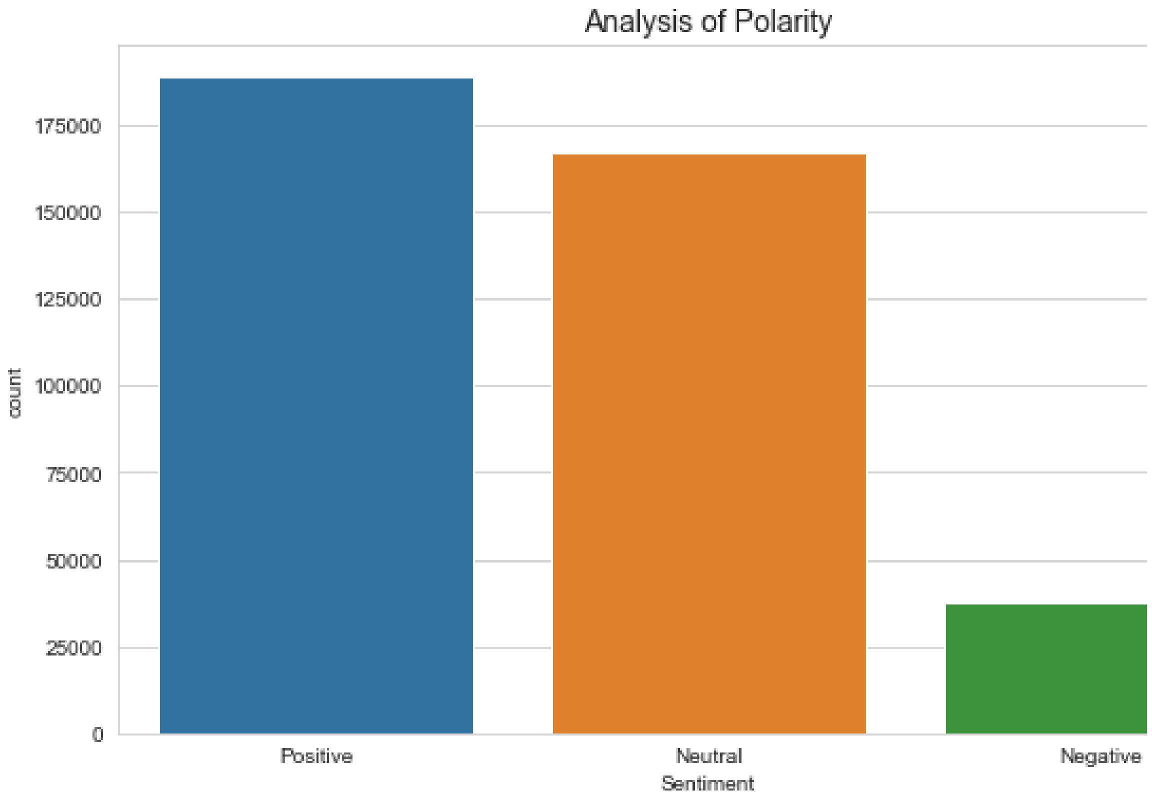
- Removing HTML Tags
- Expanding Contractions
- Removing Special Characters
- Lemmatization
- Removing Stop words

Exploratory Data Analysis (EDA)



- Most of the reviews have score 5
- very a smaller number of reviews got score 2.

Sentiment Analysis



- TextBlob
- Polarity: Polarity is float which lies in the range of [-1.0,1.0]
- Subjectivity: A float value which lies in the range of [0,1]

Sentiment Prediction using Reviews

- Logistic Regression
- The logistic regression model on the Bag-of-Words features gave an accuracy 93.8%
- Artificial Neural networks
- Artificial Neural Networks using keras API and TensorFlow managed to get a better result of 95.3% accuracy on validation data.

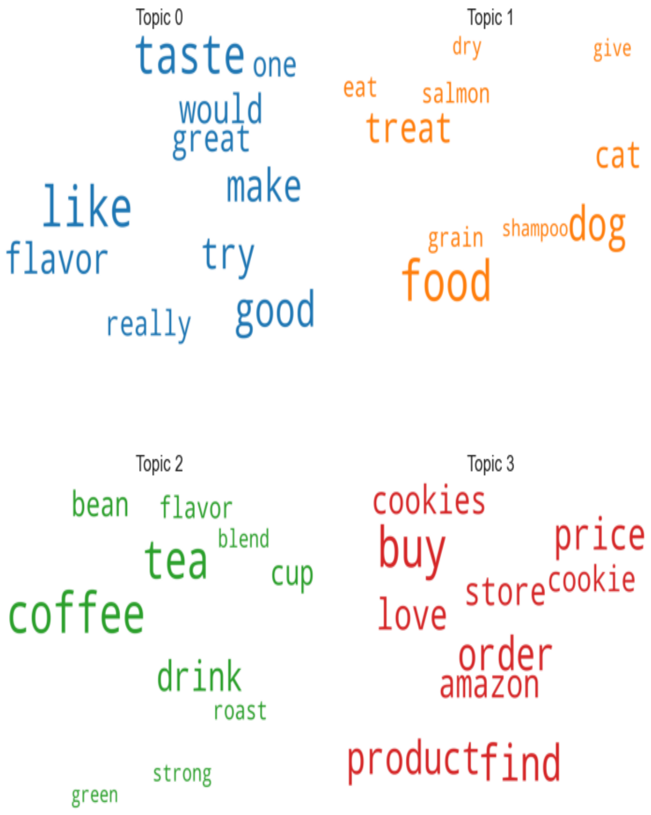
Topic Modeling

- Topic modeling helps in exploring large amounts of text data, finding clusters of words, similarity between documents, and discovering abstract topics.
- Enable consumers to quickly extract the key topics covered by the reviews without having to go through all of them.
- Latent Dirichlet Allocation (LDA) is a popular algorithm for topic modeling with excellent implementations in the Python’s Gensim package.

Steps involved

- Tokenize words and clean-up text
- Creating Bigram and Trigram Models
- Feature Engineering
- The Word2Vec Model
- Word Algebra
- Building the Topic Model
- View the topics in LDA model

Visualize the topics-keywords



Topic 1	Topic 2	Topic 3	Topic 4	
Term1	like	food	coffee	buy
Term2	taste	dog	chocolate	get
Term3	good	treat	cup	product
Term4	flavor	cat	bean	find
Term5	try	hair	milk	order
Term6	make	eat	roast	love
Term7	tea	salmon	blend	price
Term8	really	grain	dark	time
Term9	great	chew	cake	bag
Term10	one	dry	coconut	box

Text Summarization

- Extractive Summarization and Abstractive Summarization
- Text Summarization Steps

  Convert Paragraphs to sentences

  Text Preprocessing

  Tokenizing the sentences

  Find Weighted Frequency of Occurrence

  Replace Words by Weighted Frequency in Original Sentences

  Sort Sentences in Descending Order of Sum

Conclusion and Future work

- Ideally, the aim of the project was to help Amazon consumers to understand the sentiment of the review, to get consumer feedback in the form of topics covered by the reviews without having to go through all of them and Enable them to quickly extract the summary of the reviews without reading the entirely. As a future work I would like to implement the Text summarization using Artificial Neural Network.