amazon

How do people Write Reviews

Anant Gadodia - agadodi
Gabriel Francis - gfranci
Shreya S Karra - sskarra

NC STATE UNIVERSITY

amazon

Background

- Reading reviews are an essential part of the buying process as it helps buyers make informed decisions
- Amazon allows reviews to be given ‘helpful’ or ‘not helpful’ votes to allow accurate ones to stand out.
- Figuring out what makes a review helpful is not a direct task but should be achievable using good data analysis and by building a good classifier.

NC STATE UNIVERSITY

Data

Marketplace - Denoting the market in which the product was sold

Customer_id - Unique identifier for the customer

Review_id - Unique identifier for the review

productid - Unique identifier for the product

Product_parent - Unique identifier for the product given to the review for mapping

Product_title - Title of the product.

Product_category - Type of product

Star_rating - Rating given by the reviewer

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.

amazon

Proposed Methods for Text Preprocessing

Cleaning and Normalizing

- Setting all to lowercase
- Removing digits, punctuation and small words

Stopword Removal

- Removal of commonly occurring words that don't impart meaning to the text

Stemming

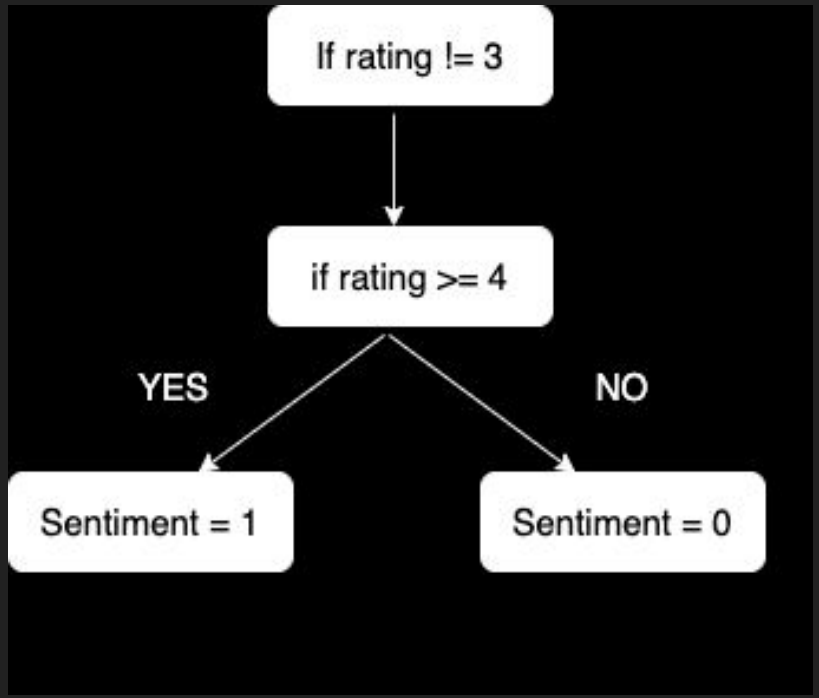
- Reducing words to their roots using a set of rules such as removing ‘ing’ from words by assuming it is a present continuous tense

Lemmatization

- Reducing words to their roots using a lexical database to understand the meaning of the word

Proposed Method for Sentiment Analysis

- First use a trained NLTK lexicon called **VADER.**
- The dataset will also be trained using a **Logistic Regression** classifier to try to improve the previous result.



amazon

Proposed Methods for Artificial Neural Network

- A neural network is useful in classification when there are hidden relationships between the features and the labels.
- The feed forward network will be trained in the classification of reviews as helpful or not.
- This can then be used to test any review without manual reading.
- The data will require additional preprocessing before it can be accepted by the network.

amazon

Experiment - Basic Analysis

Extracted 1.1M reviews

Data Cleaning:

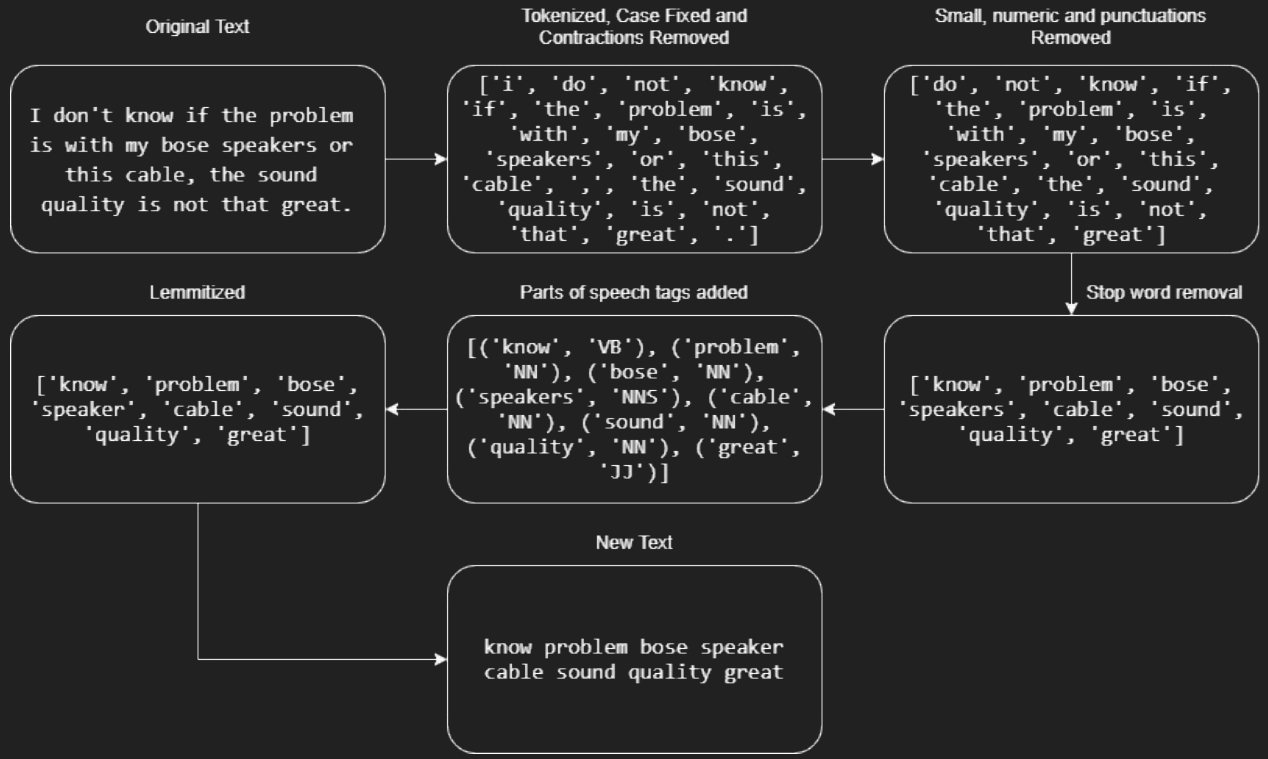
- Filling null values
- Data type correction
- Garbage value removal
- Deduplication

Data Validation:

- Checking date ranges
- Checking rating ranges
- Comparing # of reviews for ratings
- Comparing category wise metrics
- Checking rating distribution
- Checking helpfulness of reviews

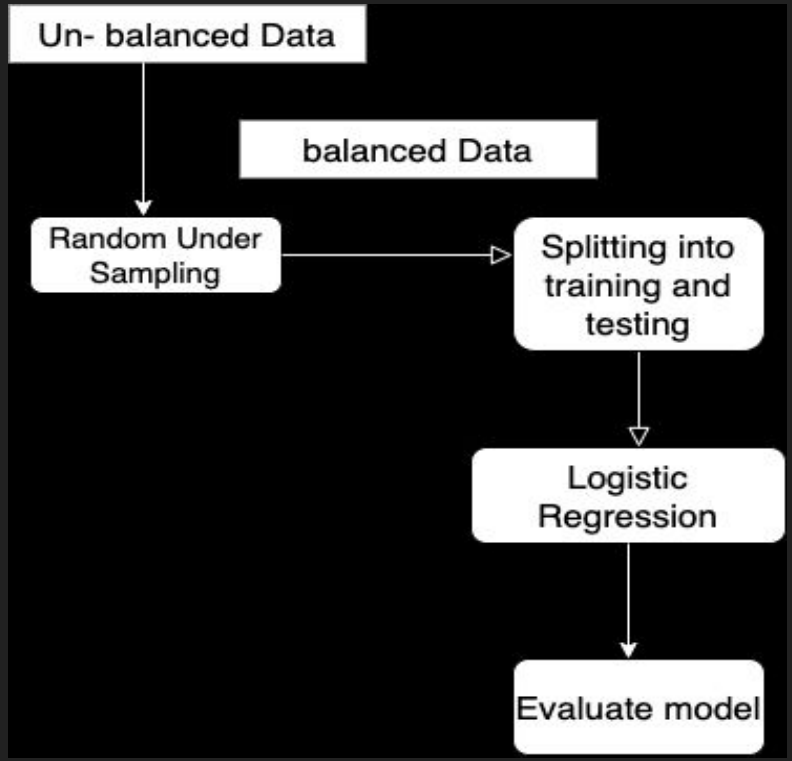
NC STATE UNIVERSITY

Experiment - Preprocessing



Experiment - Sentiment Analysis

- The logistic regression model is applied for a heavily unbalanced data which was skewed towards the number of 1s.
- The data is balanced using Random Under Sampling.
- Random Undersampling randomly selects and remove samples from the majority class till the number of samples in both the classes are equal.
- Logistic Regression model is applied to the balanced data as well.



amazon

Experiment - Artificial Neural Network

- Preprocess the data by applying one-hot encoding on categorical features and calculate the helpful_rate by dividing helpful_votes by total_votes
- Created an artificial neural network accepting Review Length and One-Hot Encoded Star Rating, Vine and Verified Purchase.
- Train the data using our training set and then validate it with our testing set.

NC STATE UNIVERSITY

Results

Basic Statistics

- Reviews analysed: 1.1M
- Categories Analysed: 17
- Average Rating: 4.14
- 63% of reviews had 5★
- Lowest star rating: Digital Software Category(3.55)

Review Analysis for Electronics

Repeated words in 1-3 star reviews

Results - Review Length vs Product Category

- After calculating the length of the review body against the product category and found some interesting results.
- The ones with the longest reviews are highly subjective items the buyer may not know the intricacies of before purchase.
- While the ones with shortest reviews can easily be previewed or predicted before purchase and thus don't need verbose reviews

Categories with Longest Median Review Length

- Books
- Digital_Ebook_Purchase
- Apparel

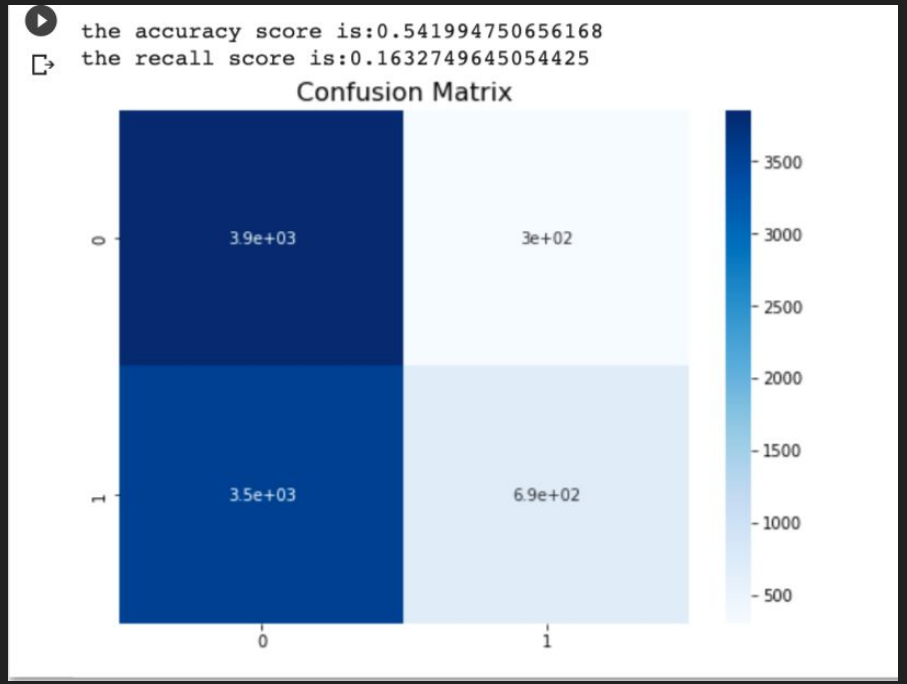
Categories with Shortest Median Review Length

- Digital_Music_Purchase
- Gift Card
- Digital_Video_Download

Results - Sentiment Analysis

- The accuracy achieved for the unbalanced model is **0.931**
- The accuracy achieved for the balanced data is **0.542**

Confusion Matrix for balanced data



amazon

Results - Helpfulness using ANN

- An Absolute Mean Error of **0.24** was achieved.
- Upon further examination it was found that the network was always giving values between 0.5 and 0.6 even with great variation of the features.

NC STATE UNIVERSITY

amazon

Conclusions

- The data still has a few hidden correlations that can be useful in our goal
- Since the accuracy for the Sentiment Analysis using Logistic Regression is 0.542, our goal for the final report is to train the dataset using Classifiers like **SVM, NB** etc and find out the model with the best accuracy.
- The artificial neural network used is essentially useless and will need to be redone to be of use.
- Other simpler and more advanced classification techniques need to be considered to get the desired result.

NC STATE UNIVERSITY