AMAZON REVIEW & RATING ANALYSIS USING NLP & NAÏVE BAYES CLASSIFIER

Presented By: Group 1

Uddhav Zambare: (002199488)

Revati Lachyan: (002197827)

Anshita Verma: (001007320)

Mani Deepak Reddy:(002728148)

AGENDA

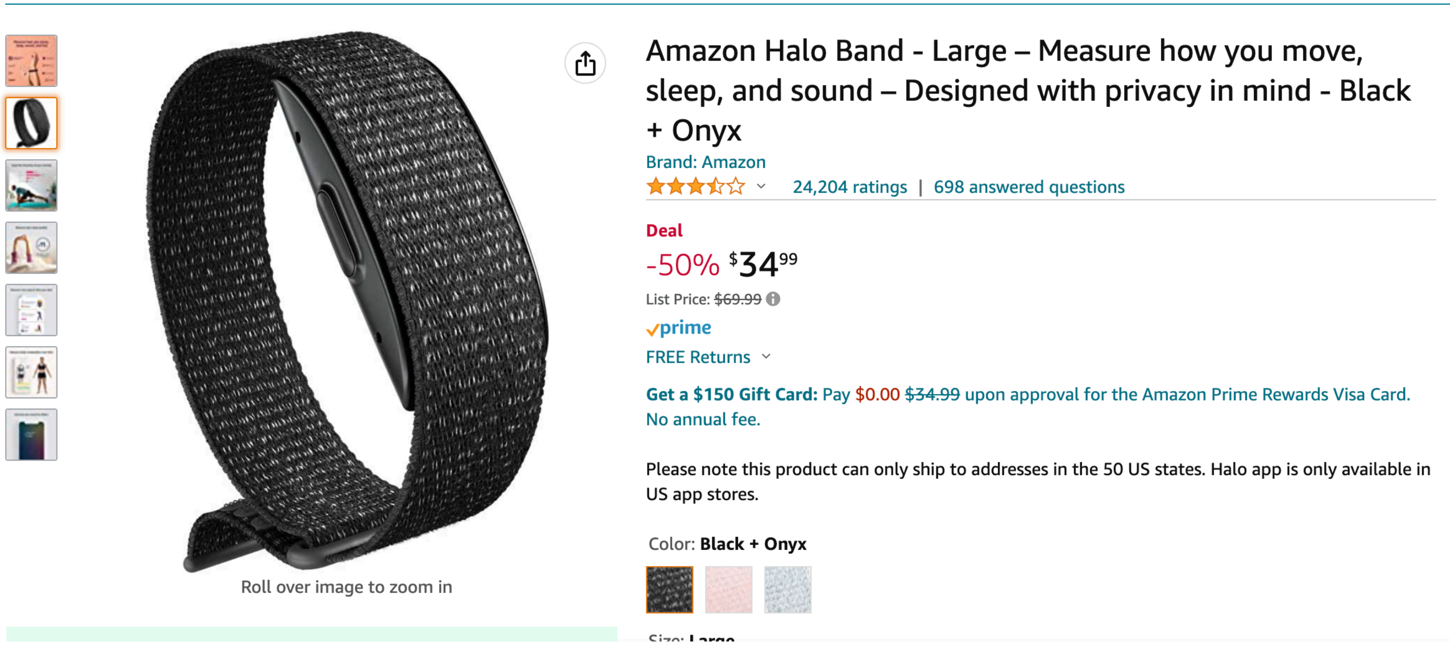
- Problem Statement
- Data Description
- Data Scraping
- Data Preprocessing
- Natural Language Processing Approach
- Word Cloud
- Conclusion

PROBLEM

Scrape about 5000 reviews on any product of your choice on Amazon along with their ratings (1-5).

Considering reviews with 1-2 as 'negative' and 4-5 as 'positive', develop an NBC classifier

DATA DESCRIPTION



Amazon Halo Band - Large – Measure how you move, sleep, and sound – Designed with privacy in mind - Black + Onyx

Brand: Amazon

★★★½☆ 24,204 ratings | 698 answered questions

Deal

-50% $34.99

List Price: ~~$69.99~~

✓prime

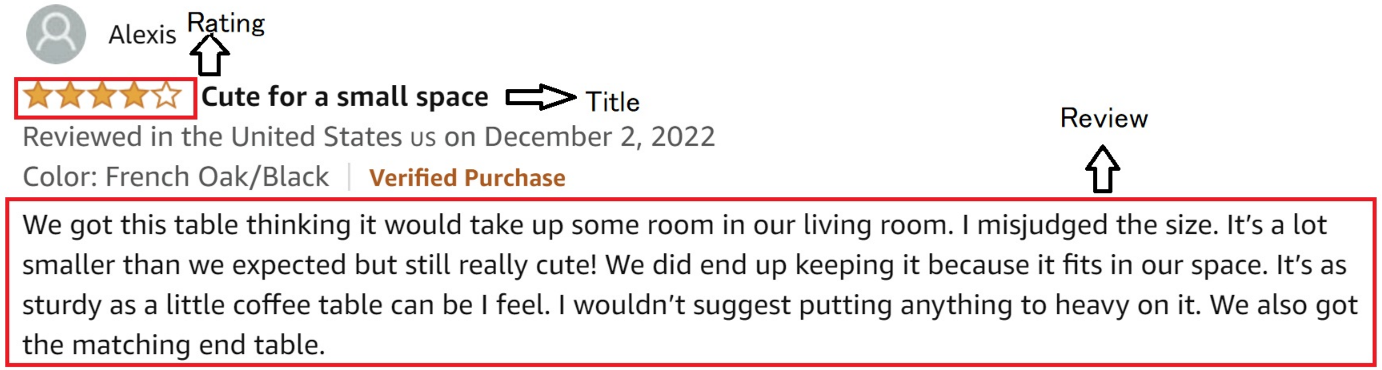
FREE Returns

Get a $150 Gift Card: Pay $0.00 ~~$34.99~~ upon approval for the Amazon Prime Rewards Visa Card. No annual fee.

Please note this product can only ship to addresses in the 50 US states. Halo app is only available in US app stores.

Color: **Black + Onyx**

Size: **Large**



Alexis

★★★★☆ **Cute for a small space**

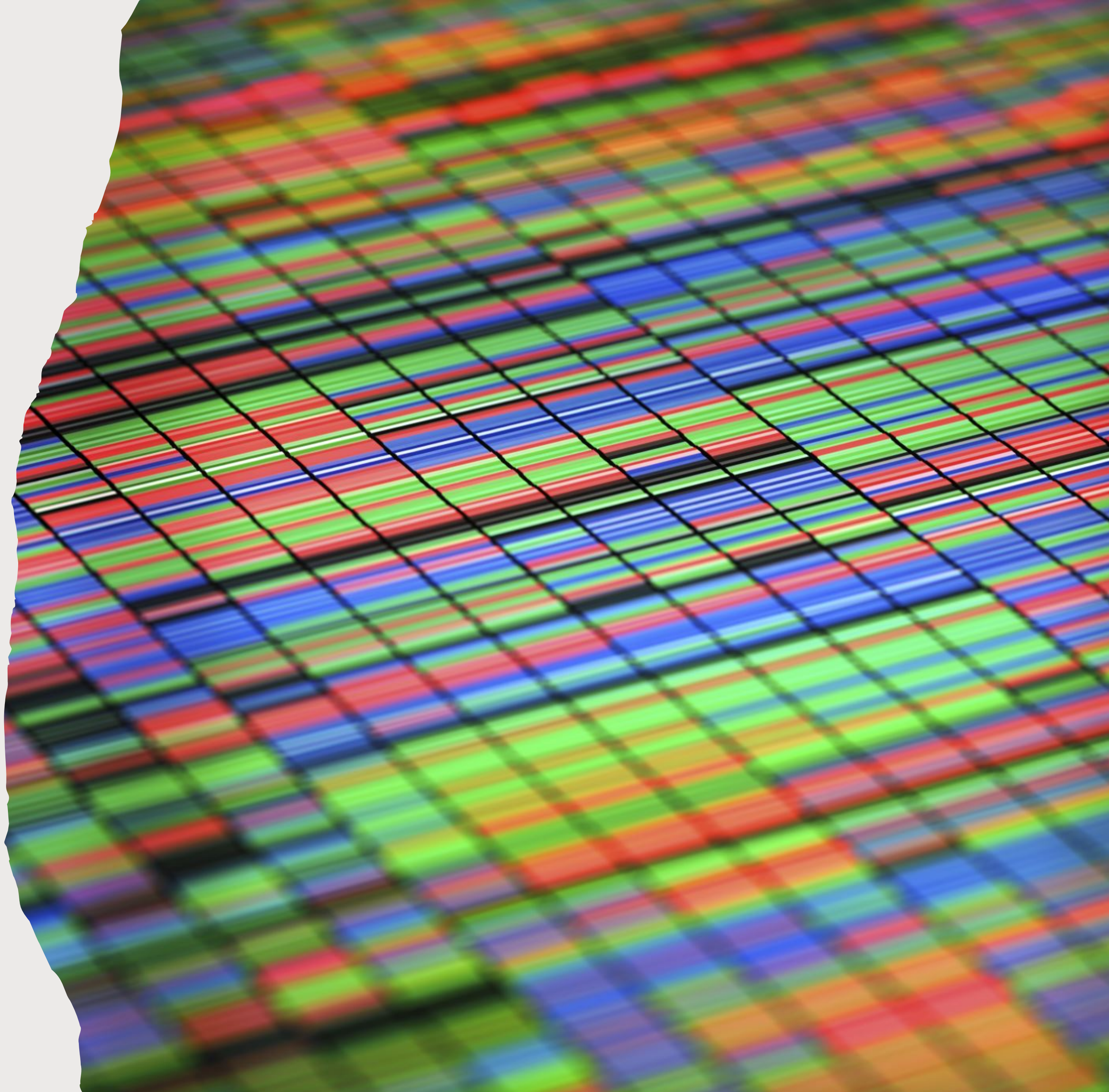
Reviewed in the United States on December 2, 2022

Color: French Oak/Black | **Verified Purchase**

We got this table thinking it would take up some room in our living room. I misjudged the size. It’s a lot smaller than we expected but still really cute! We did end up keeping it because it fits in our space. It’s as sturdy as a little coffee table can be I feel. I wouldn’t suggest putting anything to heavy on it. We also got the matching end table.

DATA SCRAPING

- API used: ScraperAPI
- Beautiful Soup: Library that allows you to efficiently and easily pull-out information from HTML



NATURAL LANGUAGE PROCESSING

- **Tokenizing:** The process of converting natural text into smaller parts known as "tokens."

  Tokenizer used: RegexpTokenizer

- **Stop Words**: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore.

  Stop Words used: stopwords from nltk.corpus

- **Lemmatizing**: Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.

  Caring-> Care

  Lemmatizer used: WordNetLemmatizer

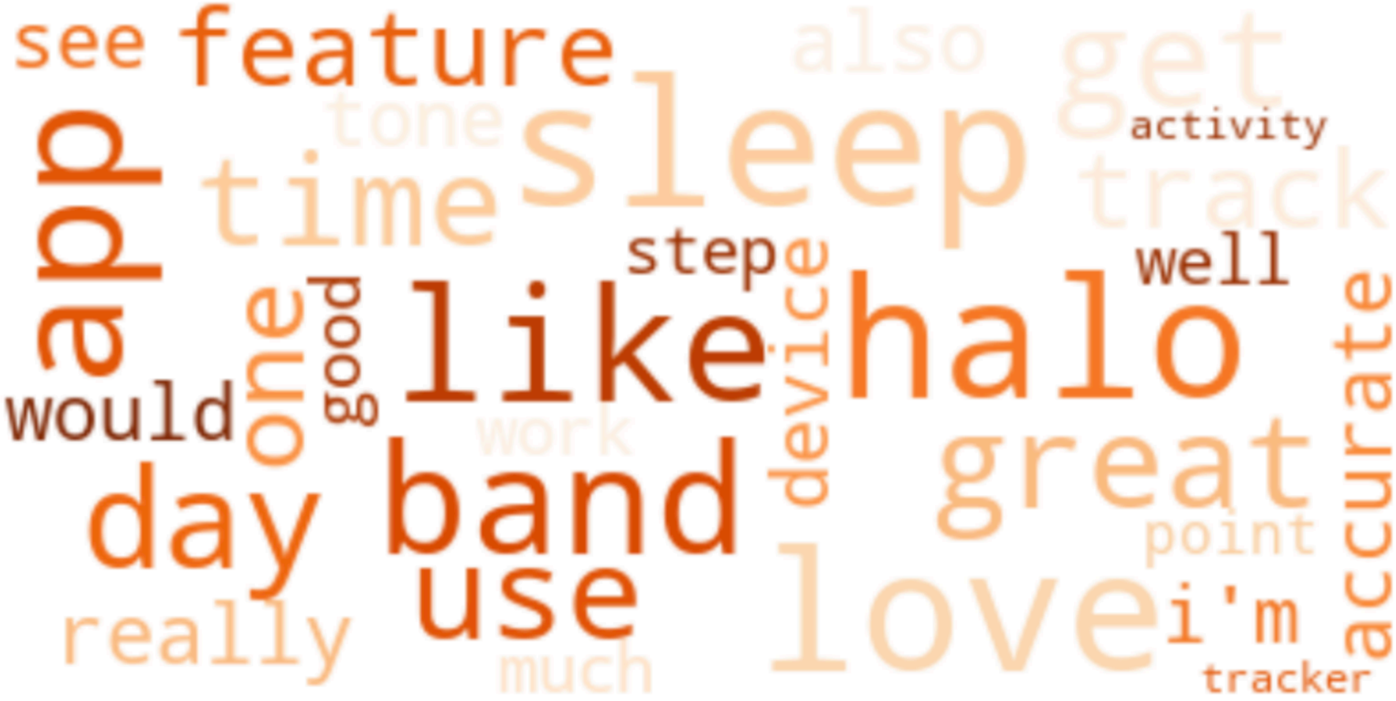
- **Removing Accents:** Converting UTF8 to ASCII
- Naïve -> Naive

NAÏVE BAYES CLASSIFICATION

- A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.
- The assumption made in this classifier are the predictors/features are independent. That the presence of one feature does not affect the other.
- **Types of Naive Bayes Classifier:**
    - **Multinomial Naive Bayes:** The features/predictors used by the classifier are the frequency of the words present in the document.
    - **Bernoulli Naive Bayes:** This is like the multinomial naive bayes, but the predictors are Boolean variables.
    - **Gaussian Naive Bayes:** When the predictors take up a continuous value and are not discrete values.
    - **Complement Naïve Bayes:** It uses statistics from the complement of each class to compute the model’s weights**.**

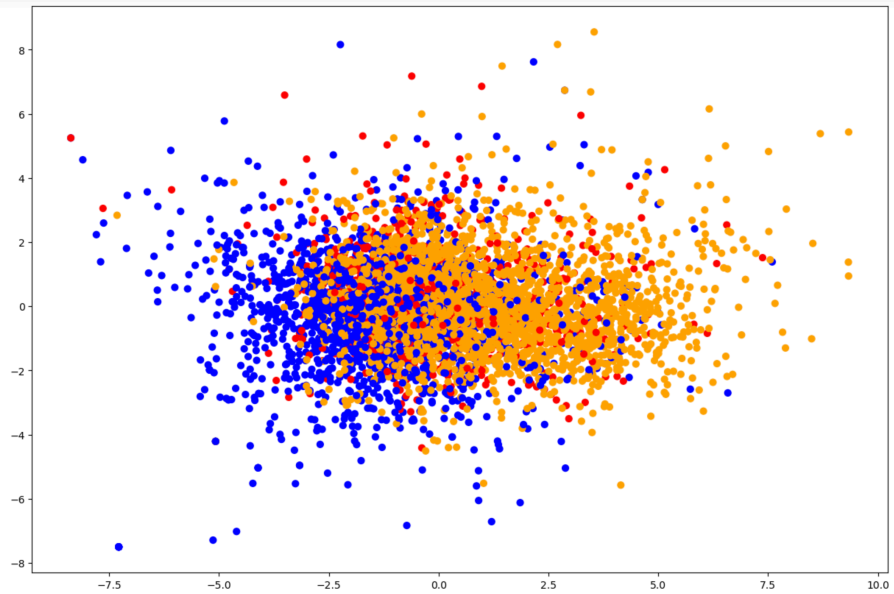
WORD CLOUDS

Word Cloud for 5-Star Ratings



FEATURE ENGINEERING

- **Bag of Words**: Convert document (a list of words) into the bag-of-words format = list of (token_id, token_count)
- **TF-IDF Model**: Assigns continuous values instead of simple integers for the token frequency. (Gensim.models.tfidfmodel)
- **Word2vec**: Embeds meaning in vectors by quantifying how often a word appears within the vicinity of a given set of other words.
- **Principal Component Analysis**: Dimensionality reduction technique



CONCLUSION

- Accuracy of Naïve Bayes Classifier with :

Algorithm	Accuracy
Multinominal Naive Bayes	88 %
Complement Naive Bayes	78 %
Bernoulli Naive Bayes	70 %
Gaussian Naive Bayes	64 %

- Accuracy using Random Forest Classifier with :

Algorithm	Accuracy
Random Forest Classifier	82 %