AMAZON BEAUTY PRODUCT REVIEWS: SENTIMENT ANALYSIS & WEB APPLICATION

STATS 418, SPRING 2024

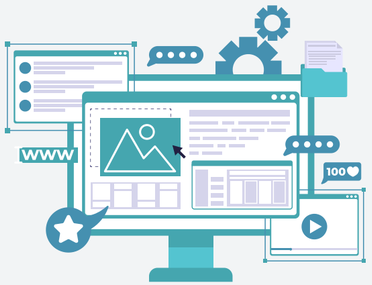
ANUM DAMANI, YUHUAN HE, DANI WU, CINDY XU

CONTENT

ABOUT THE DATASET

- **Source**: Created by Jianmo Ni, a UC San Diego PhD student currently working at Google.
- **Category Focus**: "All Beauty" Category

Column Name	*Description*
overall Rating	given by the reviewer (1 to 5 stars)
verified	Indicates if the review is from a verified purchase
reviewerID	Unique identifier for the reviewer
productID	Unique identifier for the product
reviewerName	Name/Username of the reviewer
reviewText	The full text of the review
summaryReviewText	Short summary of the review
year	The year when the review was written
processed_reviewText	The preprocessed text of the review after tokenizing and lowercasing
filtered_reviewText	The filtered review after removing stopwords



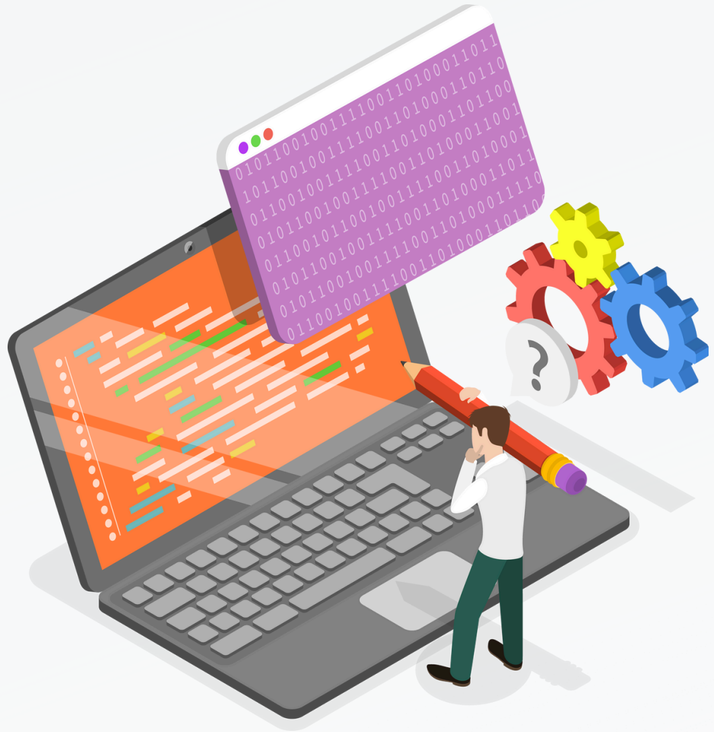
DATA CLEANING & PREPROCESSING

- **Tools**: Used Python for its powerful libraries and ease of use.
- **Transformation**: Converted JSON data to a tabular format and renamed columns for clarity.
- **Missing Values**: Verified none were present; retained only necessary columns.
- **Date Handling**: Extracted and stored only the year from review dates.
- **Text Preprocessing**: Tokenized, lowercased, removed stopwords and punctuation.
- **Stopwords**: Extended list to include beauty-specific terms.

DASHBOARD METHODS

4 Methods for Dashboard:

1. Creating the Dashboard
2. Polarity
3. Time Series
4. Word Clouds



DASHBOARD METHODS

Data Cleaning / Preprocessing & Creating Dashboard

- Data Cleaning & Preprocessing:
  - Utilized the popular ‘nltk’ library in Python
    - ‘nltk’ Library is great for tokenizing, stemming, lemmatizing, text classification, text data cleaning, etc.
- Dashboard:
  - Created using R Shiny
    - R Shiny is excellent for creating dashboards, interactive data visualizations, connecting to APIs, etc.

DASHBOARD METHODS

Polarity

- Identifies trends in customer sentiment, highlighting extremes of feedback.
- The polarity score was calculated using the `syuzhet` package in R
- The higher the polarity score, the more positive the review.
- Users can select the number of reviews to view at a time.
- Visualizes sentiment polarity to better understand customer experiences in a more general picture.
- Pinpoints areas needing improvement or attention.
- Interactive plots provide a dynamic tool for deeper sentiment analysis and insights.

DASHBOARD METHODS

Time Series

- Time series plots offer a comprehensive view of sentiment trends over time.
- Users can observe how customer sentiment regarding beauty products has evolved.
- Plots include a general sentiment overview with aggregated positive and negative sentiments across all reviews.
- Users can select specific sentiments such as:
  - Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust
- Provides deeper insights into customer reactions.
- Highlights periods of significant positive or negative feedback.

DASHBOARD METHODS

Word Clouds

- Word Clouds: visual representations of textual data where the size of each word indicates the frequency it shows up or the overall importance of that word
- We created 2 types of Word Clouds:
  1. Filtering either verified or unverified reviews of a particular year
  2. Filtering positive/negative sentiment reviews of a particular year
- Reading in each text phrase, we use the built in WordCloud function from the wordcloud package in Python to generate a visualization of the word cloud

DASHBOARD PLOTS

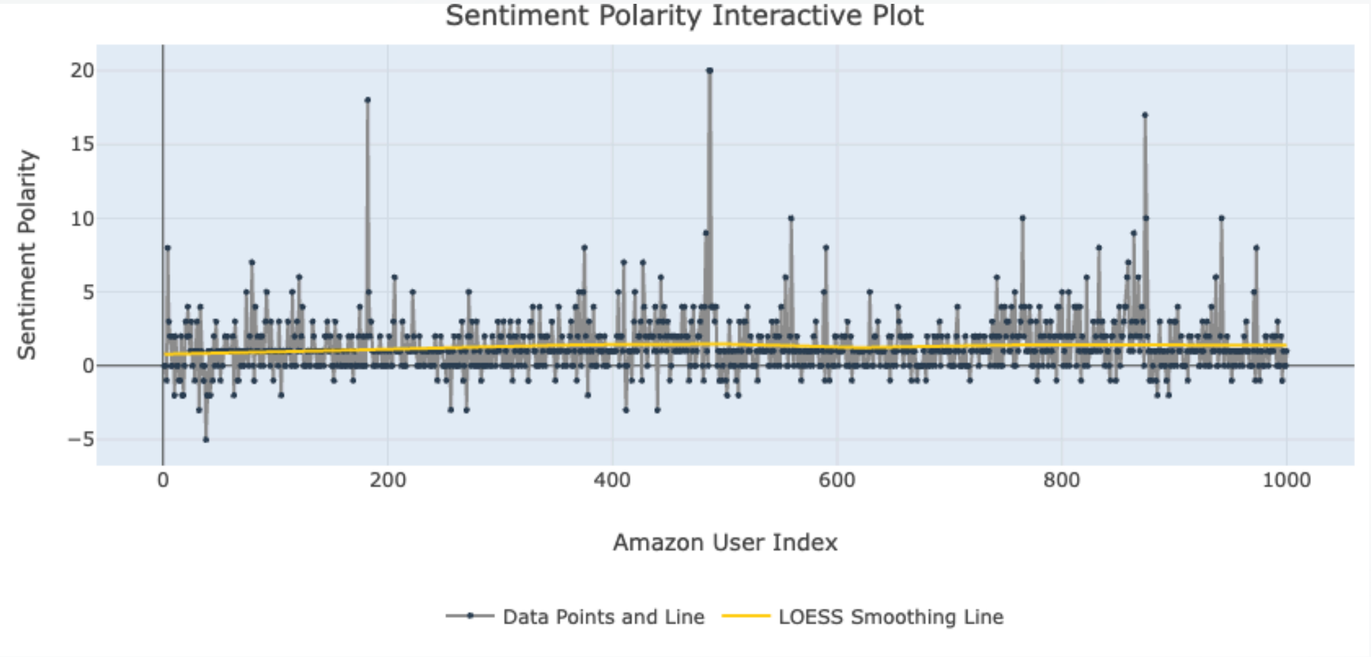
1. Polarity
2. Time Series
3. Word Clouds (User Verification)
4. Word Clouds (Sentiment Polarity)

POLARITY

Overall Sentiment Polarity Plot

- The scatter plot visualizes the sentiment polarity of each review, with a LOESS smoothing line to highlight the overall trend.
- The mean polarity score of approximately 1.447804 indicates a slight positive sentiment trend, suggesting that users generally express more positive sentiments than negative ones when reviewing products in the All Beauty category.

POLARITY PLOT EXAMPLE IN DASHBOARD

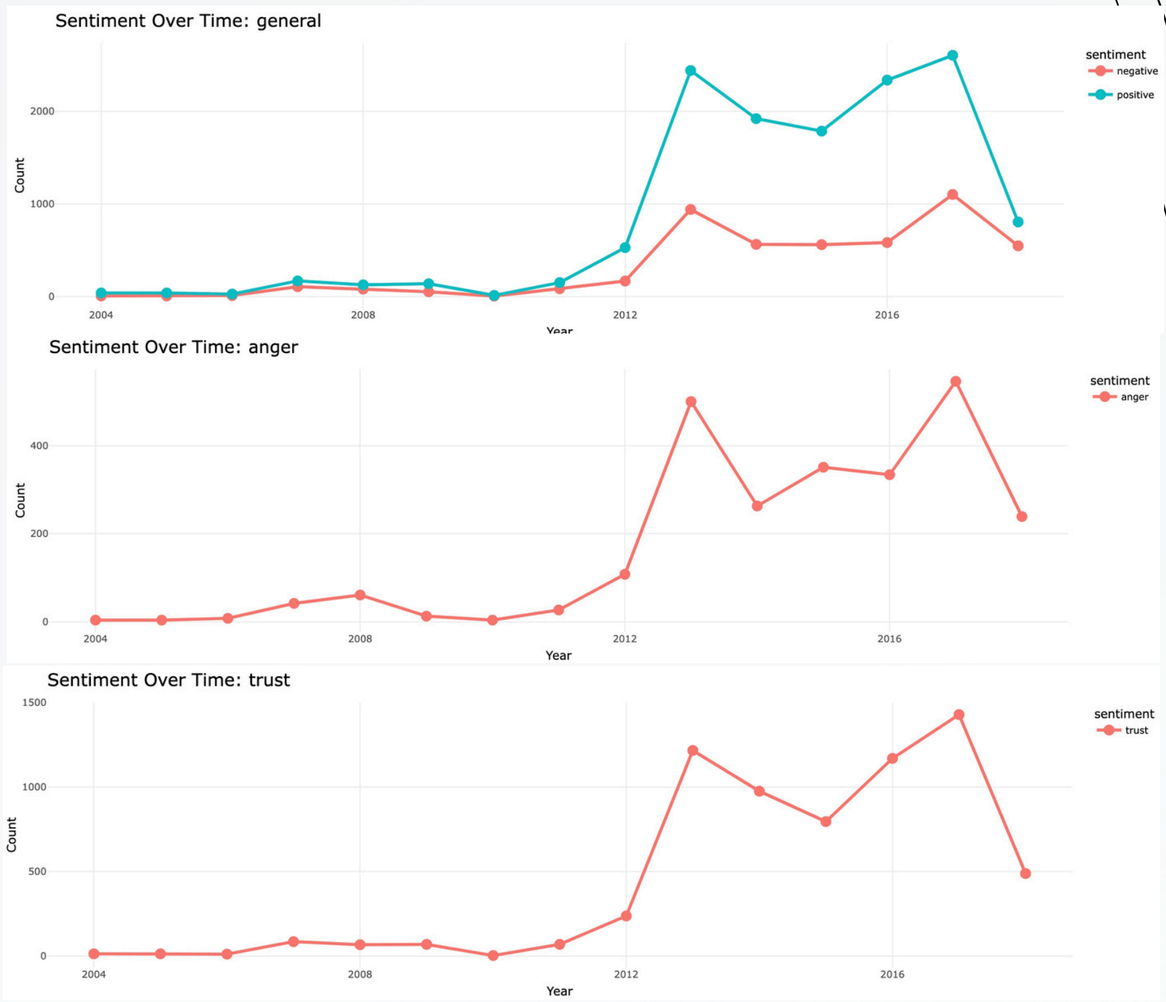


TIME SERIES

Time Series Plot for general, anger, and trust

- Noticeable spike in positive sentiment around 2013, likely due to popular product releases.
- Dips indicate periods of higher customer dissatisfaction.
- Specific sentiment trends:
  - Disgust, Anger, Sadness: Increases around 2012-2013.
  - Anticipation, Trust: Peaks around 2013, indicating growing anticipation and trust.
  - Joy: Significant increase around 2012-2013.
  - Surprise, Fear: Volatility with peaks and troughs.

TIME SERIES PLOT EXAMPLE IN DASHBOARD



WORD CLOUD

Data on Verified vs Unverified Reviews

- Word Cloud on the right is built from verified reviews of Amazon beauty products of 2018.
  - We see many words pertaining to body and hair care such as “smell,” “scalp,” “lather,” “skin,” “shampoo,” “body wash,” “clean,” and “dandruff”.
  - We also see many positive words such as “good,” “love,” “feel clean,” and “good lather” from the Amazon beauty review data with verified authenticity.

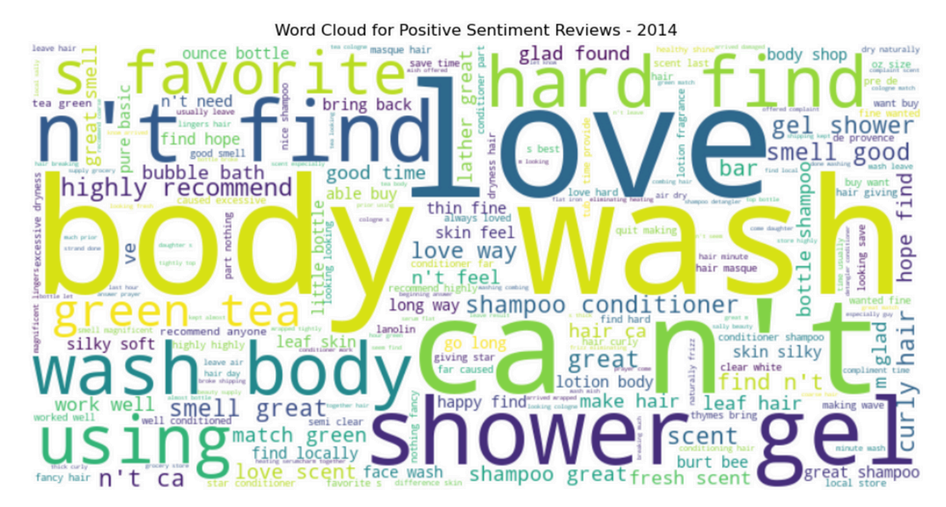
WORD CLOUD EXAMPLE IN DASHBOARD (VERIFIED, 2018)

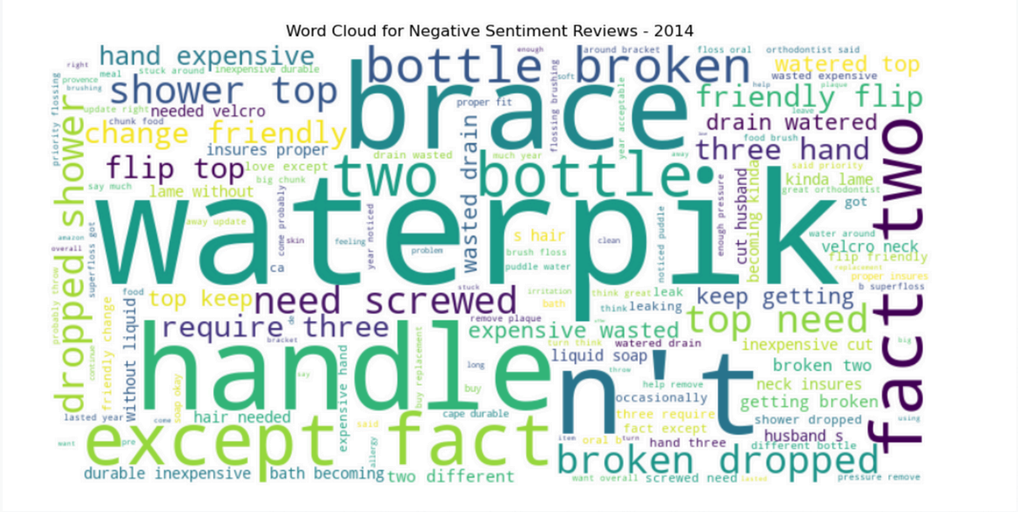
WORD CLOUD

WORD CLOUD EXAMPLES IN DASHBOARD (POSITIVE VS. NEGATIVE, 2014)

Data on Sentiment Polarity

- Wordclouds on the right are for reviews with positive and negative sentiment in 2014
  - Words with positive sentiment related to bath products
  - Words with negative sentiment related to product quality & dental products
    - Interestingly, “waterpik” was associated with positive sentiment in 2007. This suggests that waterpik quality and satisfaction decreased over time.





LIMITATIONS & FUTURE WORK

LIMITATIONS

- Analysis focuses only on beauty products
- Slang terms & ambiguous reviews may not have been classified correctly
- Dataset documentation did not explain product IDs
- We do not have access to product launch and discontinuation dates

FUTURE IMPROVEMENTS

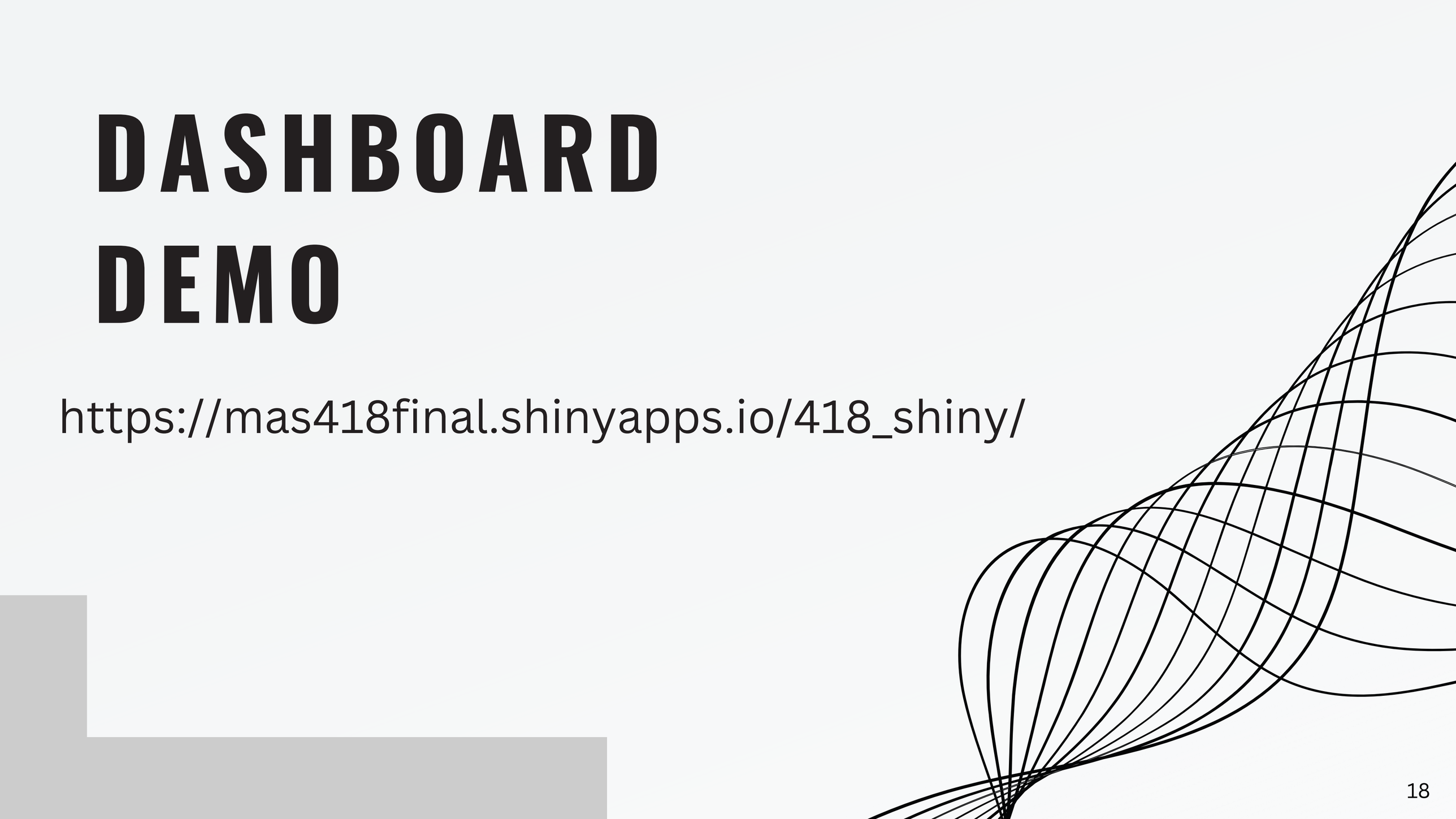
- Train a BERT model learn words bidirectionally and use to predict customer sentiment for future years
- Expand scope of project by analyzing other categories besides beauty (e.g. electronics, clothing, furniture, etc.)
- Add more advanced filtering on dashboard if we have access to product names/IDs

REFERENCES

- Ni, J. (n.d.). Amazon Review Data (2018). Amazon Review Data. https://nijianmo.github.io/amazon/
- Amazon verified purchase reviews - Amazon Customer Service. (n.d.). https://www.amazon.com/gp/help/customer/display.html?nodeId=G75XTB7MBMBTXP6W

DASHBOARD DEMO

https://mas418final.shinyapps.io/418_shiny/



THANK YOU!