Where's the Beef?

Tamar Brand-Perez, Tiffany Price, Ben Tubbs, and Jose Santos

Tiffany

Good evening everyone. Tonight we are going to answer the question of “Where’s the beef?”

Meat substitutes are emerging options for food security, food safety, and sustainability

Per Fortune Business Insights, the global population is projected to reach 9.8 billion in 2050; it is not feasible to sustain such an enormous population on animal meat alone.(https://www.fortunebusinessinsights.com/industry-reports/meat-substitutes-market-100239)

One of the most significant wake up calls is related to the pandemic. According to the NYT, farmers and ranchers who supply the nation with meat products were confronted with several crises at once: for example, during 2020, large processing plants shut down as workers fell ill as a result of COVID, and many producers were already strained by the trade war with China.

> A larger concern is around global warming; the meat business has been under growing scrutiny in recent years for its climate change consequences, with scientists and environmentalists urging Americans to eat less meat.
> (***https://www.nytimes.com/2020/04/17/climate/meat-industry-climate-impact.html***)

The plant-based meat market was estimated to be valued at $4.3 billion in 2020. (***https://www.marketsandmarkets.com/Market-Reports/plant-based-meat-market-44922705.html***) And could be a $450 billion market by 2040. (https://www.bloomberg.com/news/articles/2021-04-16/beyond-meat-bynd-impossible-foods-battle-over-future-of-fake-meat-industry) ***OVERVIEW***

The fake meat industry has recently experienced a dramatic shift in positive user sentiment. We wanted to explore this concept to ultimately provide

stakeholders with the information and tools to assist them in determining which artificial meat brand to sell.

We feel passionately about the subject and hope that our efforts will aid in increasing the awareness and sales of artificial meat products.

Tiffany
Technology used: Tableau & SQ Lite
Tools used: Git Hub, Flask, TablePlus
Language used: Python, sql, NLTK
Algorithms used: Bag of words model, Machine Learning: Naive Bayes Classifier

Now we will discuss Data Exploration & Analysis. Tamar will next spend the next few minutes sharing about our data gathering process and database.

Data

Data Source

- Amazon product data
  - Metadata
- Amazon scraped data
- Reddit scraped data

Our initial goal was to use amazon reviews to create a machine learning model that accurately predicts whether a review is positive or negative.

Our goal was to focus on fake meat products but in the reviews dataset there was no brand and product information so we added the metadata dataset hoping to combine the two and find reviews specifically about fake meat.
That did not work so we added the scraped data.
We scraped reviews from amazon as well as reviews from reddit.

Alternatives to Meat

Amazon and reddit scraped reviews were specifically of fake meat products.

Data content

Source	Amount	Star rating	Review Text	Reviewer	Date	Brand	Product	Price	asin	len
UCSD Amazon	> 1 million	★	✓	✓	✓				✓	✓
UCSD metadata	116	★				✓	✓	✓	✓	
Scraped Amazon	1073	★	✓	✓	✓	✓	✓	✓		✓
Scraped Reddit	19		✓							

This table shows an overview of the data available in each of the datasets we used:

UCSD Amazon - general groceries reviews
UCSD metadata - fake meat product and brand information supposed to be of the same dataset as the UCSD Amazon.
Scraped Amazon - review of fake meat
Scraped Reddit - review of fake meat

We wanted a dataset that had all of the columns mentioned here and most importantly, the review text as well as the product and price information because we wanted to focus only on fake meat reviews. Since the dataset from UCSD had groceries reviews but no product and brand information, we tried to merge it with the metadata that was supposed to belong to it.

The general groceries dataset from ucsd had more than 5 million reviews. After cleaning the data we had more than 1 million reviews. The metadata had a similar number initially but after choosing only the fake meat reviews, the metadata set had only 116 reviews. Then when we merged the UCSD amazon with the metadata we were left with only 12 rows. We tried many different variations of merge, join, concatinate, with an inner, outer, left joins with no success. The only common column in these datasets was the asin which is the Amazon's product number. When we looked further into the data we found that the asin in one of the datasets was made of letters and digits and the other one was digits only so this explained why the merging of these datasets did not work.

So we decided to shift to scraped data from Amazon website. We got the information we needed.
Scraped data was specifically about fake meat products. when scraping we had to input multiple url's to get the data.

The reddit reviews did not have star ratings. We scraped the reddit reviews about fake meat to show that our machine learning model could predict if a review is positive or negative therefore expanding the number of reviews that store owners could use to help them in understanding how people feel about different fake meat products and ultimately, help them decide which brands and products they may want to sell in their stores.

Data Cleaning

- Dropped columns
- Added length column
- Dropped duplicates
- 1 and 5 stars
- Trimmed the 5 stars

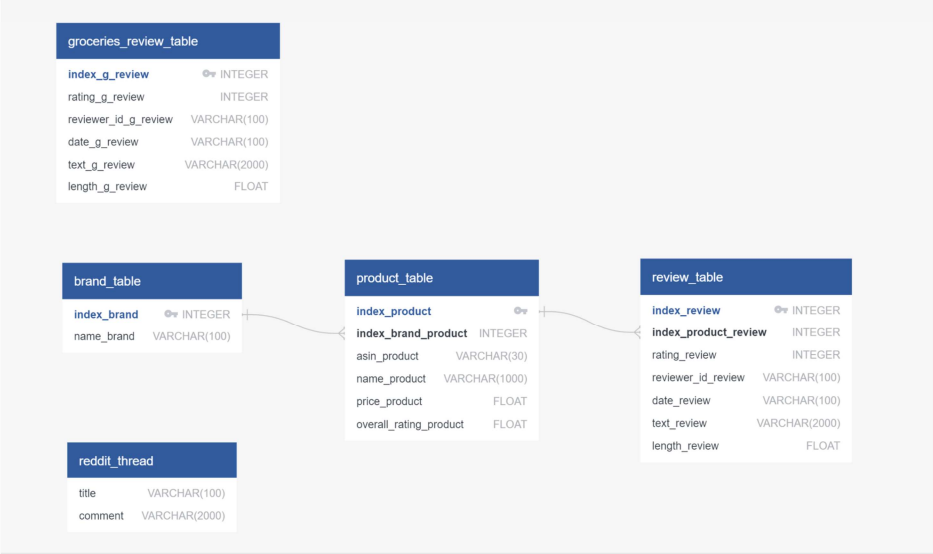
Data cleaning:

We dropped columns that were not relevant for our goal:(columns that were dropped were: verified, reviewer name, summary of the review text, unix review time,(vote, image and style - had little data in them))

We added the review length column because we were going to find out if there was a connection between the length of a review and its sentiment (We ended up not using that). This column showed the number of words in each review.

We dropped duplicate rows that is rows that all data was identical were dropped.
We narrowed the data to include only the 1 and 5 starred reviews so the difference between the positive and negative was more obvious.
We trimmed the 5 star reviews from more than 3 million reviews to 600,000 to match the number of reviews we had for 1 star to feed a balanced set into the machine learning model.

Database



Database

We used Sqlite and sqlalchemy to create the database. This was a local database that is suitable for the development stage of a project.
We interacted with the database using python, and table plus.
This ERD is showing the tables in the database. The first one is the data from UCSD amazon general groceries. (I did not include the metadata because we did not use it.) The three in the middle are from the amazon scraped data. The first one is of the brands. The second is the products. Since each brand has several products, the connection is one to many. Since each product has multiple reviews the connection between the product and the review table was also one to many. Lastly we have the reddit table with Reddit reviews.

Questions

- How do people feel about fake meat products
  - Positive or negative review?
  - Frequent words reviewers associate with a brand.
  - Price change over time?
  - Rating change over time?
- Inform grocery store owners

The questions our team hoped to answer with this data:
We wanted to collect information about fake meat and provide it to stakeholders so they can make informed decisions. We wanted to provide a better understanding of how people feel about fake meat products:

1. Did a fake meat product receive a positive or negative review?
2. What are some keywords that users use to describe specific products, which may provide insight into what characterizes a specific product in the eyes of reviewers.
3. Did the price of these fake meat product change over time?
4. Did the rating of these products change over time?

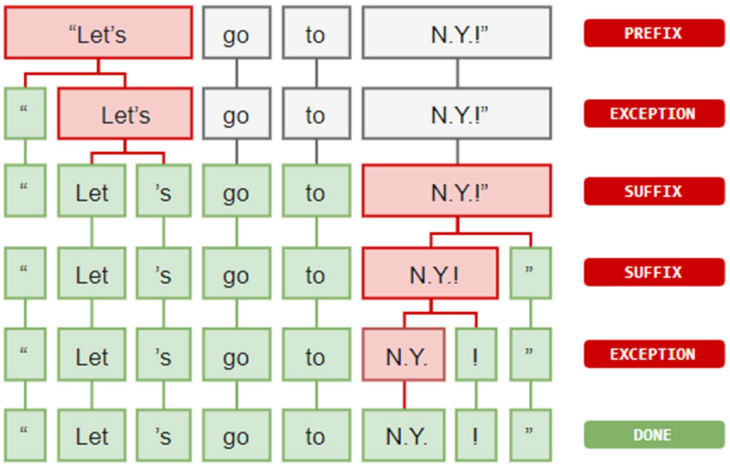
All of this information would be provided for example to groceries store owners to help them decide which products and brands they might choose to sell in their stores.

Machine Learning

We are creating a sentiment analyzer that classifies reviews as positive or negative. Our dataset is a collection of Amazon reviews that we used to train the model. Amazon reviews may contain any type of lexicon including words that are misspelled, abbreviated, capitalized, contain punctuation, etc. The first step is to normalize this lexicon so that items such as “I’m”, “Im”, “i’m” and “im” are not treated as different words.

Preprocessing and Normalization

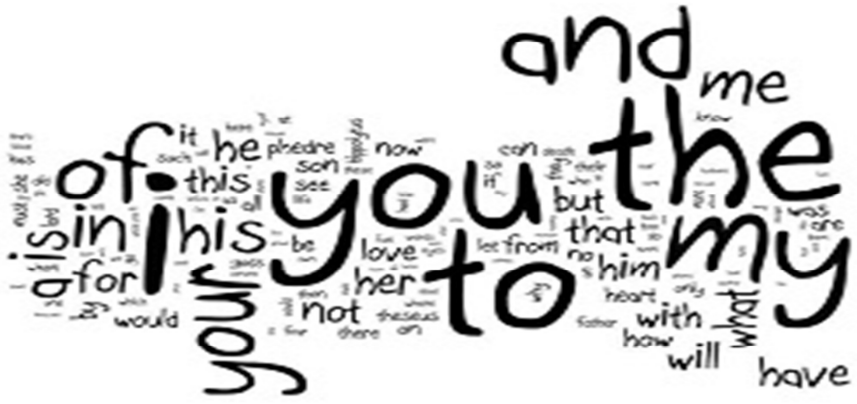
- Tokenization
- Casing
- Removing Non Alphanumerics
- Length
- Stop Words
- Lemmatization



We need to split the sentence into a list of words for processing. We could do something simple like split the sentence at each word space. Tokenizers split words in an intelligent way, so that words with periods, such as "Mr." are not treated as separate tokens.
Normalization

- **Casing** - The first step was to switch every letter of every token to lowercase, so that tokens such as "The" and "the" would not be treated as separate entries
- **Removing Non Alphanumerics** - Next all letters that were not alphanumeric were removed to prevent words such as "mr" and "mr." from being treated as separate tokens
- **Length** - All tokens that were less than length 2 were also removed
- **Stop Words** - Stop words were removed
- **Lemmatization** - All words were lemmatized for greater normalization

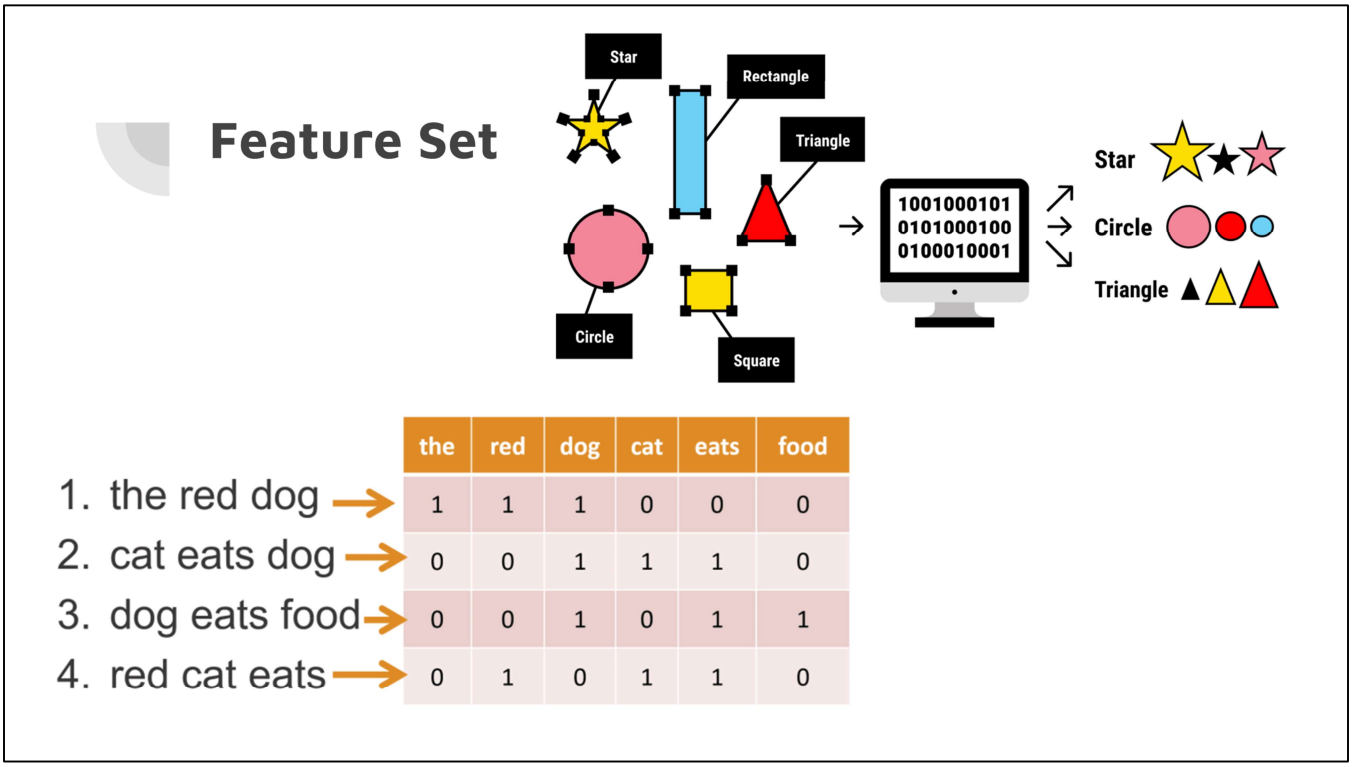
Cleaning



	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Stop words are words that do not add a lot of meaning to the text such as “the”, “it”, “as” and “about”. Stop words were removed.
The same word may come in many forms, such as: “eat”, “ate”, “eaten”, etc. All of these words can be normalized to what are called lemmas so that they are not treated as separate tokens, but as the same token.



Next, we need to create the labeled data. We used only the five star reviews as positive and the one star as negative. There were many more positive than negative reviews, so we used only 1000 of each to eliminate bias in the model. We used monograms which means each word is treated separately and word order is not taken into account. We then turned each review into a vector (a python dictionary) of the 3000 most common words in all of the reviews. This dictionary stores the presence or absence of each of the 3000 most common words in all of the reviews.

Most Informative Features

Feature	Sentiment	Certainty
Lic	Positive	17.0%
Great	Positive	12.9%
Tin	Negative	9.7%
Eat	Positive	9.7%
Received	Negative	9.4%
Dis	Negative	8.4%
Rec	Negative	8.4%
Purchased	Negative	8.3%
Receive	Negative	7.6%
Flavor	Negative	6.3%
Perfect	Positive	5.7%
Purchase	Negative	5.4%
Rip	Negative	5.0%
Tasty	Positive	5.0%
Using	Positive	5.0%

Here are the top fifteen most informative features along with whether they are positive or negative and the percentage of how telling they are.

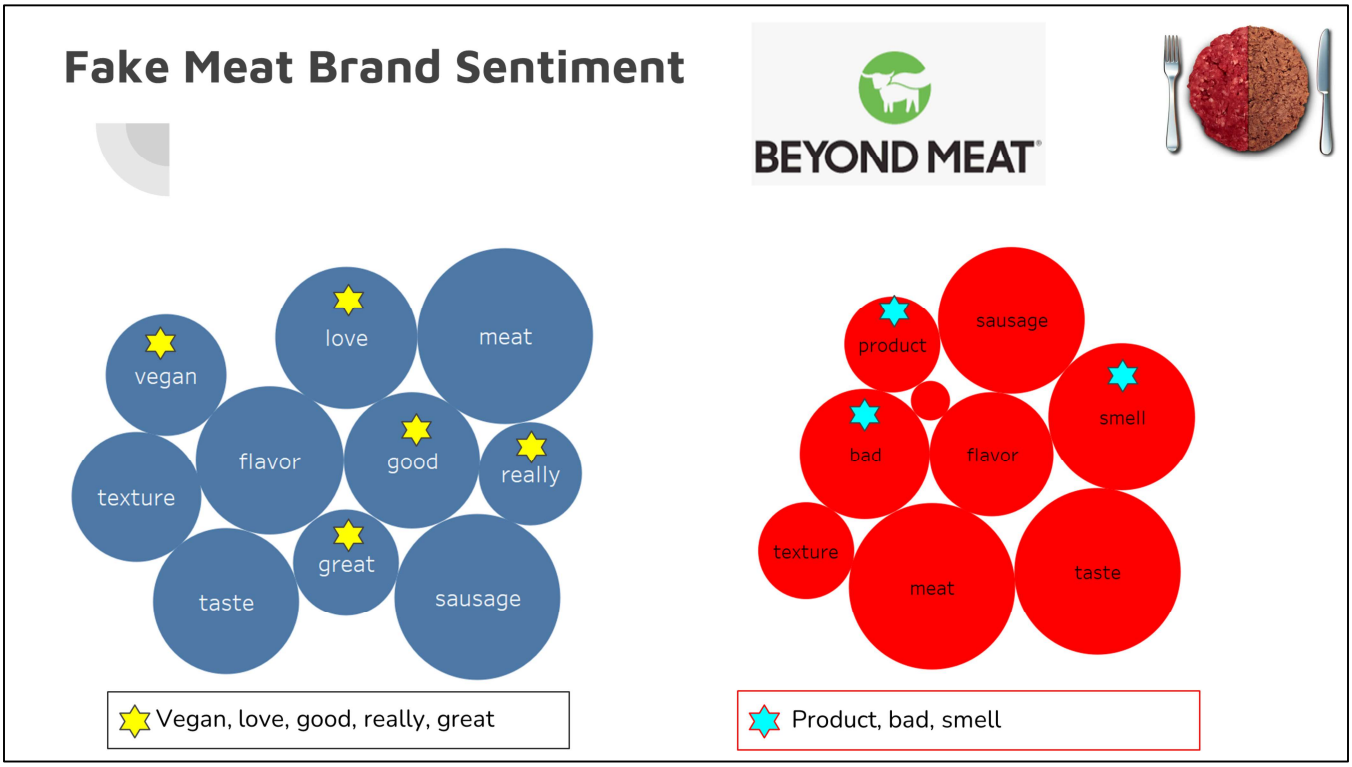
Visualizations

Tiffany - introduction
To create visualizations in Tableau, we scraped amazon for reviews and ratings of 5 fake meat brands. Using over 1,000 reviews we were able to analyze various aspects of the data. Jose will explain in detail.

Fake Meat Products Reviewed

Brand Name	Product Name
Beyond Meat	Beyond Meat Beyond Breakfast Sausage Plant-Based Breakfast Patties, Classic 7.4 oz
	Beyond Meat Beyond Sausage Plant-Based Dinner Sausage Links, Brat Original 14 oz
	Beyond Meat from PlantBased Frozen oz lb. Package, Ground Beef Substitute, 16 Ounce
Boca	Boca Original Vegan Non-GMO Soy Chik'n Veggie Nuggets (10 oz Pouch)
	Boca Original Vegan Spicy Non-GMO Soy Chik'n Veggie Patties (4 Count)
Gardein	Gardein Gluten-Free Ultimate Plant-Based Beefless Ground Crumbles, Vegan, Frozen, 13.7 oz.
	Gardein Sliced Italian Plant-Based Saus'age, Vegan, Frozen, 9 oz.
	Gardein, Burger Beefless Ultimate, 12 Ounce
Quorn	Quorn Foods Meatless Grounds, Vegetarian, Frozen, 12 Oz
	Quorn Foods Meatless Nuggets, Vegetarian, Frozen, 10.6 Oz
	Quorn, Meat-Free Meatballs, 10.6 oz (Frozen)
Tofurky	Tofurky Deli Slices Oven Roasted
	Tofurky, Deli Slices, Hickory Smoked, 5.5 oz

We scrapped reviews for fake meat products from 5 popular brands: Beyond Meat, BOCA, gardein, Quorn and Tofurky.
On average we scrapped reviews from 2 products for each brand: 3 products for Beyond Meat, 2 products for BOCA, 3 products for gardein, 3 products for Quorn and 2 products from the Tofurky brand.



Two word clouds to show the 10 most frequent words used in positive (blue circles) and negative reviews (red circles) of Beyond Meat products.
Area of circle is proportional to the frequency that the word appeared in the reviews (normalized per sentiment).
Stars indicate words that are only used in one sentiment review (positive or negative).

This view allows a stakeholder to associate keywords that define a brand with positive and negative reviews. By doing so, the stake holder is afforded a comparison between positive and negative "traits" per brand (as shown here) or positive/negative traits for a brand comparison (alternative view).

Starred words indicate words with high frequency in a sentiment review and that do not show up in high frequency on the opposite sentiment review.

Keyword frequency by brand in positive reviews

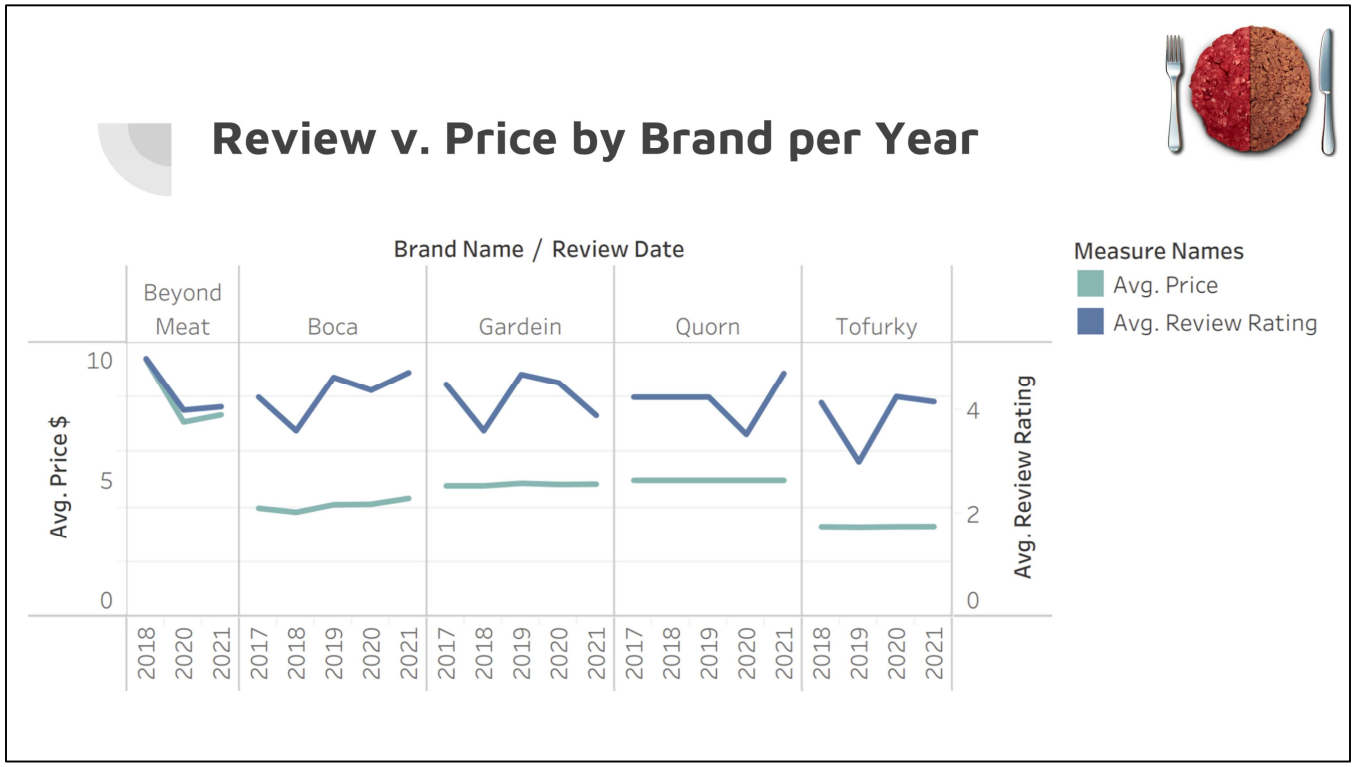
Brand Name	flavor	good	great	love	meat	real	really	taste	texture	vegan
Beyond Meat	36%	25%	22%	33%	29%	27%	22%	28%	33%	21%
Boca	6%	12%	11%	6%	2%	10%	9%	9%	1%	10%
Gardein	32%	23%	25%	24%	35%	33%	18%	27%	37%	28%
Quorn	11%	16%	18%	19%	20%	22%	29%	24%	18%	10%
Tofurky	15%	23%	24%	18%	14%	8%	21%	12%	11%	32%

Info on how often a keyword (top 10 from all reviews) is used in a review of a product of a brand - brand to brand comparison. E.g. out of total number of times that “flavor” was used in positive reviews for all brands (100%), 36% of those instances were found in reviews of Beyond meat products.

<u>This view allows the store owner to perceive how each reviewer associates a keyword in a POSITIVE review with a certain brand.</u>

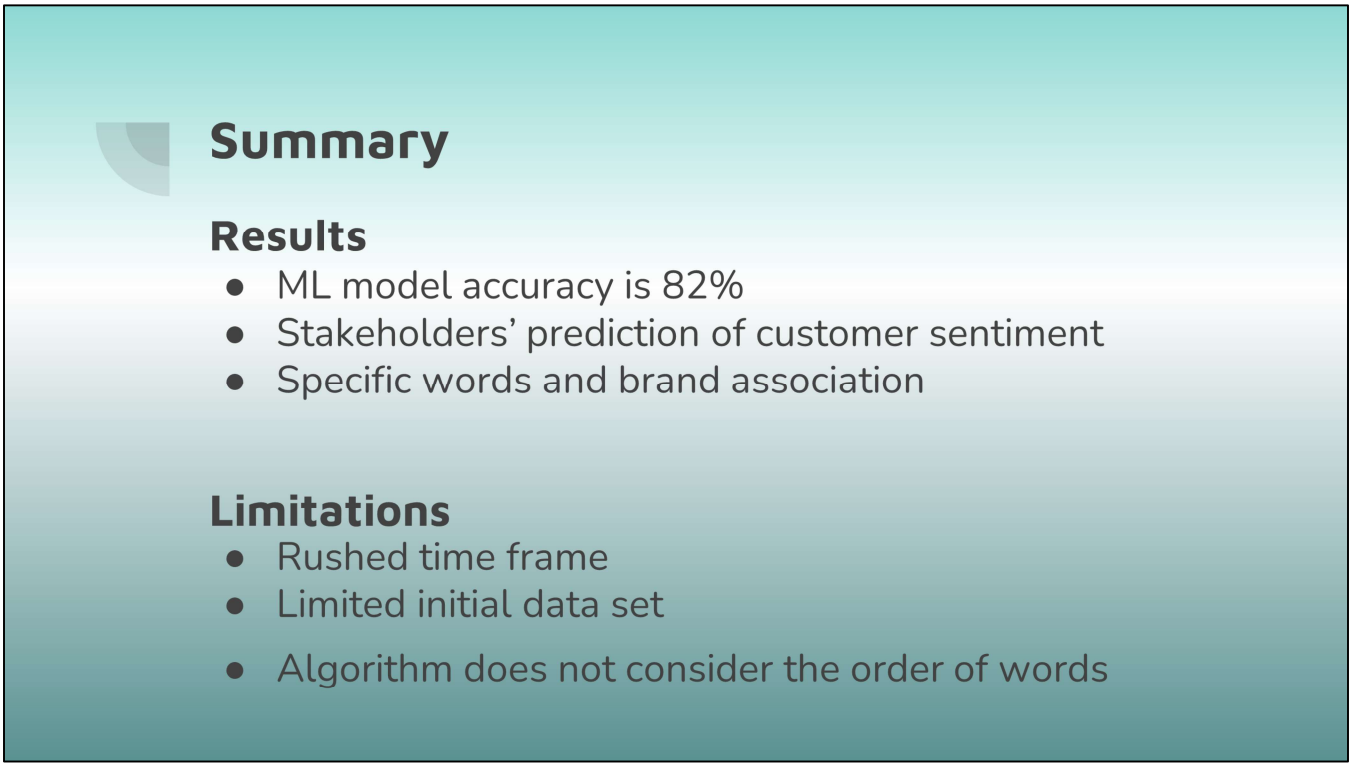
The bordered cells indicate the brand for which that particular word was used with the highest frequency.

Cells are colored from light blue (low frequency) to dark blue (high frequency) according to percent observed down a column.



A plot of the evolution of average price and average review rating for each brand.

This plot allows the stakeholder to perceive 1) evolution of consumer sentiment for each brand, and 2) how the evolution in sentiment correlates with evolution in price.
For example, for the “Beyond Meat” products reviewed, the average price appears to be well correlated with the average review rating since when the average price appears to change at the same rate and in the same direction as the average review rating. For the other brands there appears to be no correlation between price and consumer sentiment evolution.



Tiffany

RESULTS

In conclusion, our hope is to facilitate increased sales of artificial meat products by way of empowering store owners in their decision making process of which products/brands to sell.

Our ML model predicts the sentiment of user entered reviews with approximately 82% accuracy. We have seen a range of 80-84% depending on which data is used to train the model.

Also, through use of our interactive website, stakeholders have the ability to understand customer sentiment upon testing reviews.

Finally, from the word clouds and the table analysis of the reviews, the stakeholder will be able to associate specific product description words with the respective brand. This allows for a deeper understanding of how customers choose to comment about a specific product and brand.

LIMITATIONS

- Our time frame was rushed; as were learning about the limitations of our data set, we were already using it

Since our initial data set was limited, we were not necessarily focused on the best data set available. We operated with what we had and made the most of it due to time and resource constraints, including but not limited to,
computer capacity and also eventually being blocked by Amazon during our web scraping.

Finally, the algorithm does not consider the order

of words (ie food tastes great v great tasting food)

Future Analysis

- Interactive visualizations
- Larger data set
- Location data

Ways to Improve Project

- Search for a more accurate algorithm
- Have algorithm consider word order
- Host website on server

Tiffany

FURTHER ANALYSIS

If we were to continue future analysis, we would:

- Include more interactive visualizations for the end user
- Scrape in a larger data set, as available

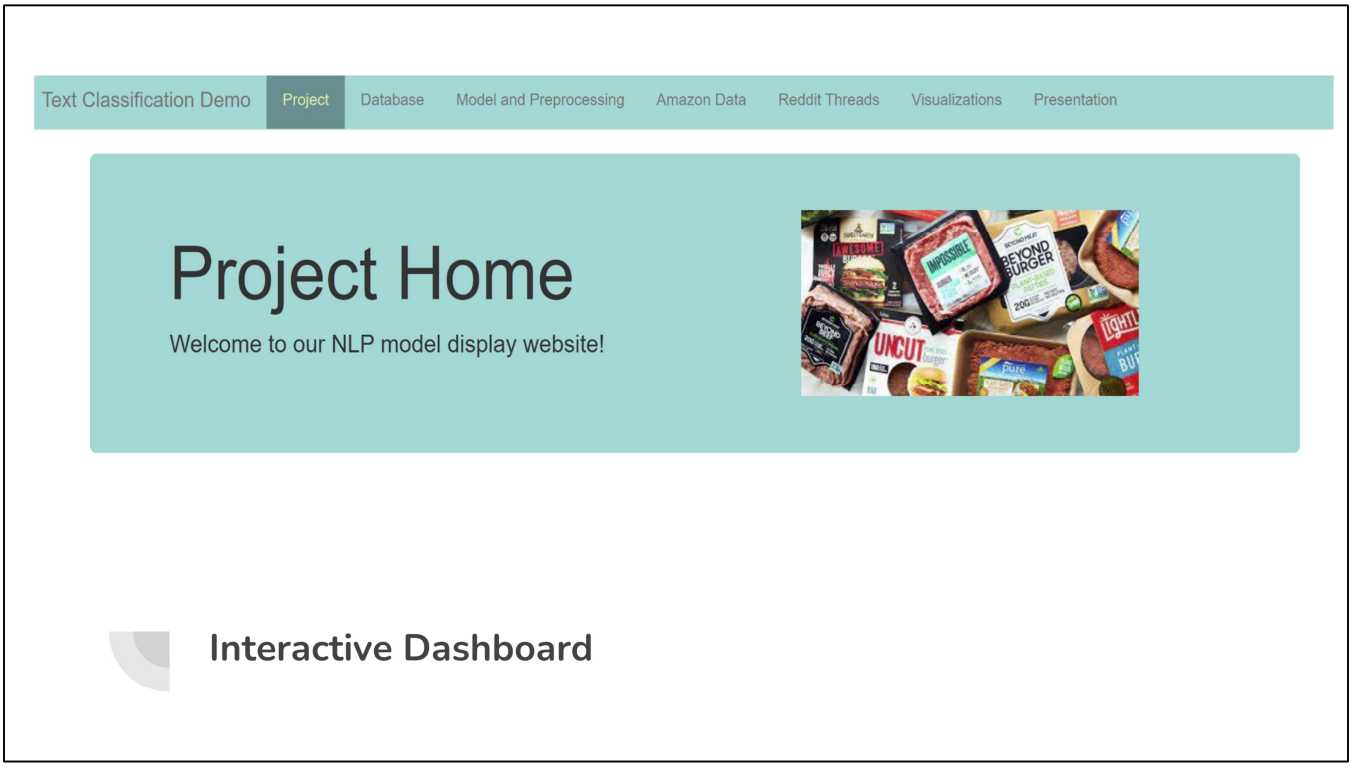
Pull in location data of reviews to further enable stakeholders' insight into consumer sentiment by location. This would allow for more market focused guidance.

WAYS TO IMPROVE PROJECT

- We would search for a more accurate algorithm
- We would have our algorithm consider word order
- Finally, we would host our website on a server

Thank you and now Ben will present our website and the informative interactive features.

Interactive Website Demo



Ben
Explanation of website/dashboard and interactivity features to predict if a review will be classified as a positive or negative review.