\*\*Author\*\*:[Russell Pihlstrom](mailto:rgpihlstrom@yahoo.com)

## Using Product Reviews & Topic Modeling to Drive Innovation!

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/Intro.png" height="300"/>

## Results 60% Coherence, 6 Topics!

Project Summary: This project is my Capstone assignment for the Flatiron School. The target audience is a fictional team of Marketers and Product Developers looking to ascertain the key product attributes consumer as saying are critical for a great analog plant-based meat product. The findings of this project, outlined below, would then be used to formulate a positioning strategy, and set of product formulas that would deliver against the positive attributes highlighted in the reviews along with addressing the negative and or missing attributes mentioned in the captured reviews.

Technical Summary: The project uses an LDA Topic Modeling algorithm to discover the latent topics within the set of approximately 3000 products reviews scraped from the sites of the most popular retailers along with twitter. After scraping and cleaning the data I created two sets of data, one set contained positive reviews the other set contained negative reviews. Finally, I created topics for each respective dataset: Positive topics and Negative Topics. A summary of these topics is immediately below, and a more detailed outline is further below. The algorithm used was able to achieve coherence scores of <b><ins>50%</ins></b> and <b><ins>60%</ins></b>, positive reviews and negative reviews, respectively. Summary of Topics:

<p>

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/SnapshotTopics2.png" height="300"/>

## Business Problem

Grocery store sales for plant-based meats are up 47% vs. year ago. Research suggests a rise in acceptance of plant-based burgers by mainstream consumers. A large portion of the growth in the plant-based category is coming from the sales of analog products (those that mimic real meat). The two most dominate brands & products in the analog plant-based meat space are: Impossible and Beyond.

<p>

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/PlantBasedSales.png" height="300"/>

<br>

<table cellspacing="0" cellpadding="0">

<tr>

<td width="70%" border="0"> The CEO of Sunrise Health Foods is looking to enter the growing category with a new plant-based burger product. However, given the high rate of failure for new product launches, coupled with the high costs for entry he/she is looking to create a product that is uniquely better than the industry leaders Impossible & Beyond. To create a new and unique product, new and unique insights are needed. The CEO has hired me to review the vast number of product reviews captured from various on-line retailers/ sources to mine for insights that could be used by the team to ensure a successful new product launch. The findings will help the Marketing & R&D teams better understand the consumers needs, wants, and desires along with how current solutions are failing to deliver.</td>

<td><img src="https://github.com/rgpihlstrom/Project5/blob/main/images/SunriseCEO.png" height="300"/></td>

</tr>

</table>

#### Business Questions Driving Model Development.

1. Identify the key sentiments & topics mentioned from product reviews.

2. Examples of good & bad product attributes associated with incumbents. These will fuel product development efforts.

<br>

#### The Data:

The data used for this project was scraped from the below sites in the listed quantities: The dataset contains approximately 3000 reviews.

<br>

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/OverviewOfTheData1.png" height="300"/>

#### Model Development Methods:

This project uses the Crisp DM methodology to generate and optimize used models. Crisp DM requires blending of business strategy, available data, and modeling techniques dictated by business drivers. Model development is and was very iterative. I began by doing secondary research around the basic business drivers of the industry, and gaining a better understanding on the top selling products and brands of plant-based meats (thanks to my friends a IRI :) ). Along with the project requirements noted above, the following additional factors were considered during the modeling process:

- \*\*1. Data Imbalance\*\* Early in the development process it was obvious that I would be dealing with an imbalanced set of data (more information/ rows of data on positive reviews vs. negative reviews). To offset this problem, I split the data into two sets: positive and negative. This step ensured the negative topics/ attributes would be surfaced without being clouded by the vast majority of positive reviews.

- \*\*2. Selection of Unsupervised Topic Modeling Techniques\*\* Initially I tried several different types of topic modeling algorithms, ranging from Mallets LDA, HDP, Multicore LDA, LSA, other. Ultimately, I decided to use <b><ins>Genism’s Multicore LDA model</ins></b> for this project. Genism’s Multicore LDA topic model outperformed the other available LDA algorithms by continuously producing more coherent topics with higher coherence scores during my iterative approach to modeling.

- \*\*3. Technical Challenges - LDA & Reviews with Few Words:\*\* Discovering the optimal number of topics, finding The "right" words for each topic, and finding topics that represented a portion of each doc/ review is/was a very iterative process. LDA Topic modeling is a technique that is considered a "soft" classifier and interpreting results are part-art-part-science. It requires the user to discover and enter the “right” number of K topics into the model. Experts suggest the higher the coherence score the better. However, the max coherence score does not always translate to the most interpretable/ translatable topics. To overcome this challenge users are required to use a blend of domain knowledge, coherence score, coherence overlap, and topic interpretability to select the “right” number of topics. During this project I achieved coherence scores that ranged from .2 - .72. Ultimately, I chose hyper-parameters that created models that reached coherence scores greater than .45 and then selected the number of topics per dataset based on interpretability. I also cross referenced my number of selected topics per dataset against a Kmeans elbow graph to ensure alignment. Lastly, I used pyLDAvis.gensim to ensure separation of topics. Below are the results from both the positive and negative datasets associated with each topic model. Below is a snapshot of the pyLDAvis.gensim visual used to aid in selecting the “right” number of topics for each dataset.

#### Model Results:

After several iterations, the below topic models were created <b><ins>50% Coherence - Positive Reviews</ins></b>, <b><ins>60% Coherence - Negative Reviews</ins></b>, These results were achieved using the <b><ins>LDA Multicore in gensim</ins></b>. It is important to remember these results were achieved with a focus on interpretability not maximizing coherence scores.

<br/>

<h5><b><u>Topics Generated From Positive Reviews</u></b></h5>

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/TopicModelingPositive1.png" height="400" />

<h5><b><u>Explanation</u></b></h5>

As stated above, Topic Modeling is part-art-part-science. Using topic models requires the user to augment the output from the algorithm to ensure usability. The algorithm produces clusters equal to K - the topic number specified when running. The output is a numerically labeled (0-K) set of clusters, along with the words that have the highest probability to be generated from each cluster. It is up to the user to name/ summarize each cluster into topics by reviewing the key words, along with the reviews that most epitomize each cluster. By studying the keywords and reviewing the reviews most associated with each cluster the user “transforms” each number cluster into themes/ topics.

The below topics were an output of this process – part-art-part-statistics 😊 .

<strong> - Topic 0</strong> = <b><ins>NEW FAMILY FAVORITE</ins></b> Per the pyLDAvis graph above, topic 0 (as labeled initially by the algorithm) was by far the most dominate topic in all the data. Keywords included: real meat, great taste, great substitute for beef, most epitomizing reviews showed raving amazement for the similarity to beef and acceptance by the family.<br/>

<strong> - Topic 1</strong> = <b><ins>WORTH THE SACRIFICE</ins></b> Keywords included: textures, reception of price, filling, and being impressed. Also shown are concerns on long/ odd ingredient lists aka (chemicals). However, given this set of topics is only drawn from positive reviews, and upon reviewing reviews associated with this topic, consumers believe the benefits outweigh these concerns<br/>

<strong> - Topic 2</strong> = <b><ins>RIGHT PROPOSITION</ins></b> Keywords and associated reviews included: size of the product and number of items contained in the package along with all the benefits associated with being animal and ecofriendly. <br/>

<strong> - Topic 3</strong> = <b><ins>GUILT FREE CONVENIENCE </ins></b> Keywords and associated reviews included: ease of use, recipe versatility and the health benefits associated with avoiding the hormones associated with non-organic beef.

<br/>

<h5><b><u>Topics Generated From Negative Reviews</u></b></h5>

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/TopicModelingNegative1.png" width="775" />

<h5><b><u>Explanation</u></b></h5>

<strong> - Topic 0</strong> = <b><ins>BUYER BEWARE</ins></b> Per the pyvis graph above this was by far the most dominate topic in the negative set of reviews. The reviews here were blasting the horrible smells, resemblance to wet petfood and its very large gap between the textures offered vs. real beef.

<br/>

<strong> - Topic 1</strong> = <b><ins>NOT WORTH THE SACRIFICE</ins></b> as the reviews most associated with this topic and the dominate words correlated with long list of ingredients, ingredients that cannot be pronounced (aka chemicals) and the high-cost relative to beef.

<br/>

<br/>

#### Example Reviews From Scrapped Sources Associated with Key words:

<img src="https://github.com/rgpihlstrom/Project5/blob/main/images/ExampleReviews.png" />

## Summary of Product Areas for Marketing & R&D Teams To Consider:

Overall incumbents’ offerings are well received, meeting most of the needs around:

</br>

- Positive Topics/ Attributes:<b><ins> Taste, Texture, Convenience, Price, Eco</ins></b>

<br/>

<br/>

Opportunities for Innovation, while maintaining positive attributes notes above include:

<br/>

- Negative Topics/Attributes:<b><ins> SMELL, APPEARANC, PRICING, INGREDIENTS </ins></b>

<br/>

<br/>

## Next Steps - MORE DATA More Data & More Data

- \*\*MORE DATA - Retailers Should Motivate More On-line Consumer Reviews\*\* - Quantity & Diversity of Thought: One of the challenges of this project was amassing enough data to obtain the quantity of data required to create robust models. In order to encourage this type of analysis in the future and across several categories more reviews would be helpful, also pushing consumer to generate organic thoughts for targeted product attributes, dislikes, etc.

- \*\*MORE DATA - Developing More Expendables Models\*\* - More Sources of Data: Per the above obtaining data is a challenge. Additionally, obtaining data that is also rated is even more of a challenge. In order to combat this problem I could develop an algorithm that would take the ratings from reviews that allow the user to enter a ratings (amazon) and using that data to project a rating on to reviews from social media sites that do not allow the user to enter a rating (Twitter,TikTok, FB). By creating a more extendable model will allow for the use of more sources of data, and ultimately increases the quantity. This is especially true for procuring negative slanted comments, this was particularly evident in this project.

- \*\*MORE DATA - Leveraging Video Content to Create Reviews\*\* - While scraping data for this project, sites such as twitter, resulted in leaving many great reviews out of the analysis. I encountered several tweets that contained the words "Beyond Burger Review" or "Impossible Burger Review" that were not really usable reviews but instead were really just links to youtube/ other videos containing the actual reviews. If there was a technology to summarize videos down to a review would be invaluable.

See the full analysis in the [Jupyter Notebooks](folder) or review our <a href="https://github.com/rgpihlstrom/Project5/blob/main/Presentation.pdf">Presentation</a>.

For additional info, contact me here: [ Russell Pihlstrom](mailto:rgpihlstrom@yahoo.com)