Team 1. Project writeup

Lavrentev Vladimir, Lobanov Ilya, Petukhov Andrey, Khaibrakhmanov Arthur

Description of a problem

Project Goal

Is to leverage the provided dataset for identifying distinctive characteristics of brands compared to their competitors within the same category, aiming to inform and enhance their future marketing strategies. This will include analyzing consumer perceptions and ratings to uncover unique brand attributes and competitive advantages, thus enabling more targeted and effective marketing initiatives.

Motivation

Our project might be helpful to many companies in high-level competitor and market analysis, identifying the strengths and weaknesses of their brand based on unique associations, and utilizing these strengths to effectively manage processes within their marketing campaign development.

Classification of a problem

Our project provides a solution of predictive problem using machine learning and pretraind ML models.

Description of data

1. Brands' identification attributes (brand ID and brand name, we will work with the latter);
2. Associative arrays given by the respondents – the associated descriptions of our brands;
3. Numerical attributes contain scores for each brand (the attractiveness and cheerfulness of the brand, etc.) - these attributes are more convenient to work with. This offers a comprehensive view of brand perception.

	collage_id	brandName	brandId	description	charming_no	cheerful_no	confident_no	contemporary_no	cool_no	corporate_no	...
0	14	7UP	4	some refreshment pictures here to make feel ha...	3	2	4	4	4	3	...
1	47	7UP	4	It's relaxing, refreshing and cool.	4	5	4	4	5	3	...
2	87	7UP	4	It represents summer, fun, the pleasure of dri...	4	5	4	3	4	1	...
3	186	7UP	4	7UP reminds me of summer in the suburbs. It's ...	1	5	4	4	5	3	...
4	259	7UP	4	7UPis crisp, refreshing, and light. I've used ...	2	4	1	2	3	3	...

Methods and models

1 Sentiment Analysis using NLP (Natural Language Processing)

Using pretrained models to quantify textual descriptions associated with each brand collage to get a sentiment score for each brand. This will reflect the spectrum of emotional responses ranging from negative to positive. We used model called "RoBERTa". RoBERTa was created by researchers at Facebook AI, and it's designed to achieve better performance and accuracy in a wide range of NLP tasks. What is special about this model:

RoBERTa is based on the transformer architecture, which uses attention mechanisms to understand the context of a word in relation to all other words in a sentence, rather than just the words immediately surrounding it. This bidirectional context is a key feature of these models, allowing them to understand language in a more nuanced way.

This model was trained on a much larger and more diverse dataset compared to BERT (a groundbreaking model developed by Google). It uses not just the BookCorpus and English Wikipedia (the datasets BERT was trained on) but also additional data sources like CC-News, OpenWebText, and Stories. This expansion of training data impacts the model quality greatly.

RoBERTa is a more robust and finely tuned version of BERT, benefiting from a larger and more varied training dataset, longer and more intensive training, and several key improvements in its training methodology. These enhancements allow it to better understand and process human language, making it a powerful tool for a wide range of NLP applications. The output of RoBERTa can be a simple classification (e.g., positive or negative), a score that represents the strength of the sentiment, or a detailed breakdown by aspects or emotions. In our case we choose to use a score, which will be helpful for the next step of our project – regression of this score on numerical attributes.

Multiple Linear Regression model:

i Model Equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- y is the dependent variable (sentiment score). - $x_1, x_2, \ldots, x_k$ are the independent variables (respondents' replies).
- β_0 is the constant term.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for each independent variable.
- ϵ represents the error term.

ii Coefficient Estimation: In multiple linear regression, the coefficients are estimated using the matrix form of the least squares method:
- The coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ is estimated as:

$$\boldsymbol{\beta} = (X^\top X)^{-1} X^\top Y$$

- Here, X is the matrix of independent variables (including a column of ones for the intercept), and Y is the vector of the dependent variable.

iii Matrix Notation: The model can be expressed in matrix notation as:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

2 Distance Measurement within Brand Categories

Measure the distance between the brands based on their TF-IDF encoded word attributes. We have calculated the central brands for each category (the centroids) to determine the most representative brand.

Methods:

Cosine Similarity (and Distance)

Cosine distance is the most common metric, when we talk about vectros' similarity. For two vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ cosine similarity measures the cosine of the angle between two vectors. The cosine distance is 1 minus the cosine similarity. For $\mathbf{x_i}$ and $\mathbf{x_j}$:

$$\text{Similarity}(\mathbf{x_i}, \mathbf{x_j}) = \frac{\mathbf{x_i} \cdot \mathbf{x_j}}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|}$$

$$\text{Distance}(\mathbf{x_i}, \mathbf{x_j}) = 1 - \text{Similarity}(\mathbf{x_i}, \mathbf{x_j})$$

Feature space compression and Searching for Centroids (using t-SNE model)

Let $\mathcal{X} \subset \mathbb{R}^{n \times p}$ be the initial feature space p, where n is the number of observations. We compress this space to $\mathbb{R}^{n \times 2}$.

t-SNE works the following way:

1. Pick a pair (x_i, x_j) from $\mathcal{X}$.
2. Calculate probabilities p_{ij}, which reflect the similarity between x_i and x_j in the initial space.
3. Project these objects to a lower dimension space, obtaining new points $\mathcal{Y} \subset \mathbb{R}^{n \times 2}$.
4. Calculate the probabilities q_{ij} for the new feature space.
5. Minimize the divergence between distributions p_{ij} and q_{ij} with GD.

Mathematically:

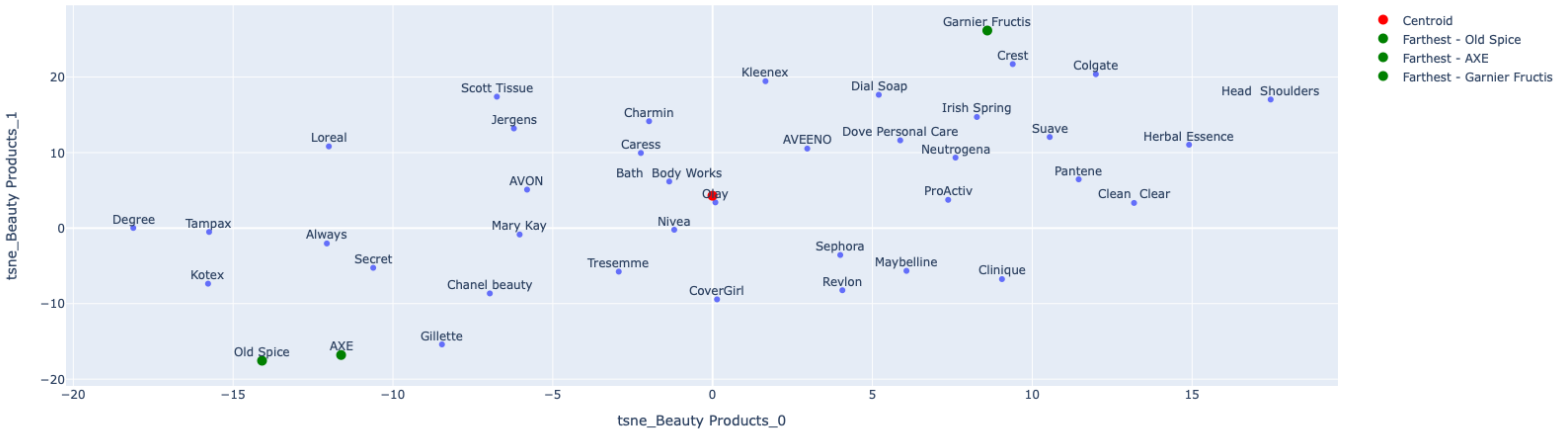
$$p_{ij} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2 / 2\sigma^2)},$$

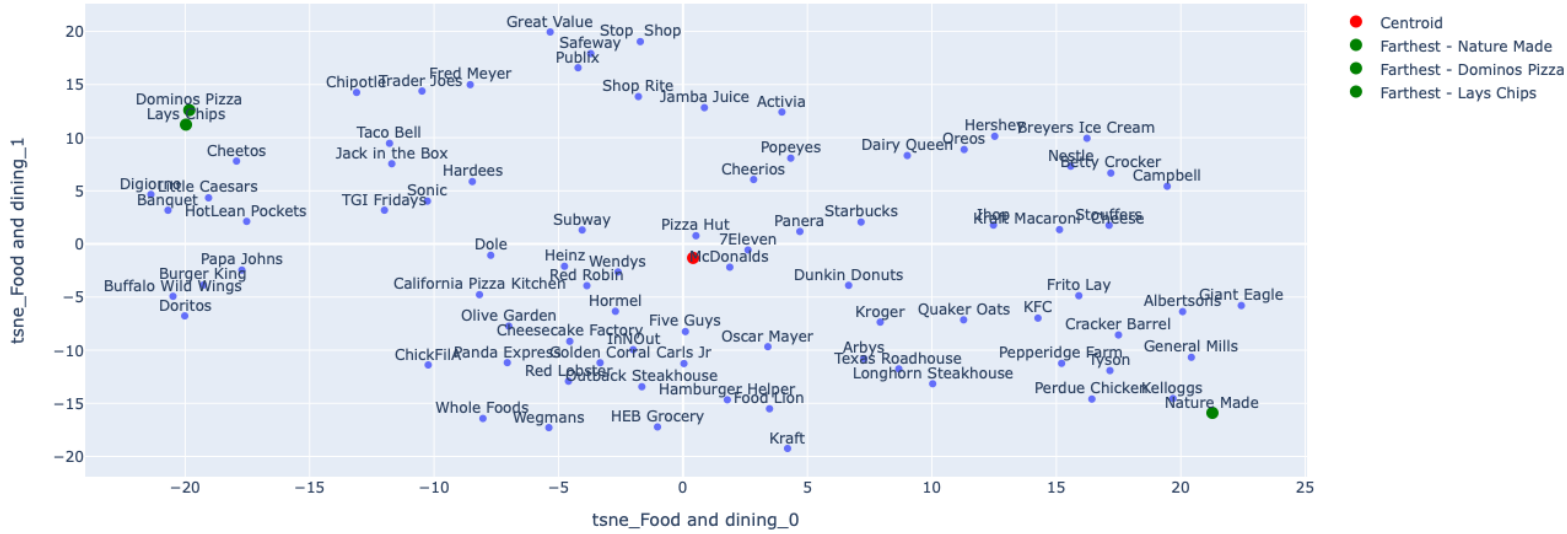
$$q_{ij} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq l} \exp(-||y_k - y_l||^2)},$$

where $y_i, y_j \in \mathcal{Y}$ are the projections of x_i, x_j in the new space.

After t-SNE we get the new feature space $\mathcal{Y}$, and now we find the centroids. Centroid C is the mass center of $\mathcal{Y}$:

$$C = \frac{1}{n} \sum_{i=1}^{n} y_i.$$





3 Peripheral Brand Analysis

The peripheral brands of category i are determined by the following formula:

$$b_1^i = \text{argmax}_j\{(y_1^{i,j} - c_1^i)^2 + (y_2^{i,j} - c_2^i)^2\}$$

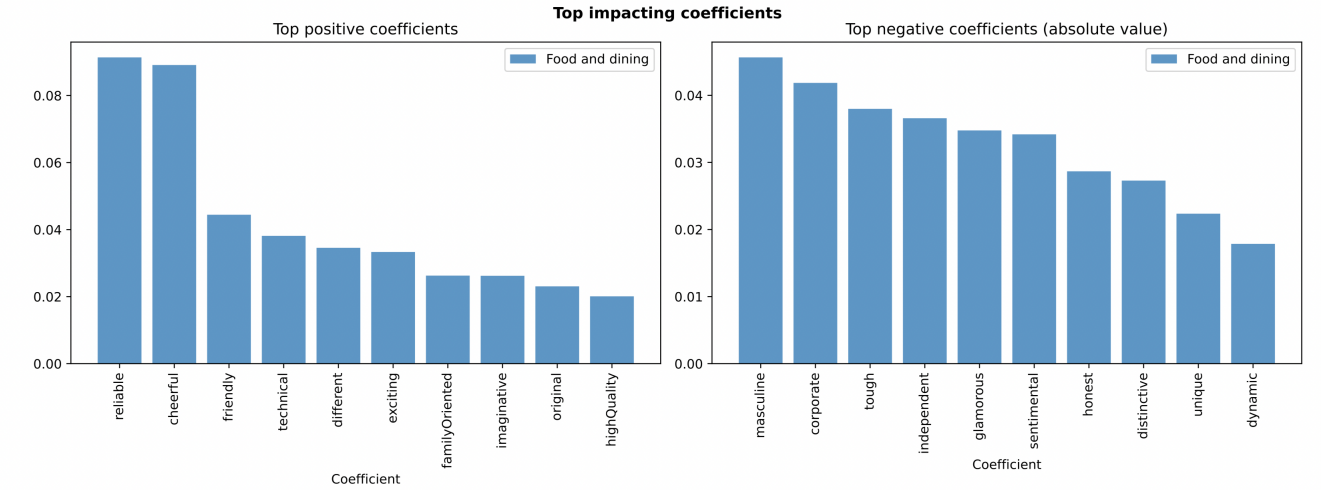
Where i is the category, $C^i = (c_1^i, c_2^i)$ is the centroid and we calculate the farthest point from it. Points b_2^i, b_3^i are determined in a similar fashion.

After determining the centroids for each category, we focused on the analysis of brands on the circumference. These brands are supposed to posses distinctive textual descriptions and set them apart from their competitors. This distinction is crucial, as it highlights unique associations and characteristics that are not commonly shared with other brands in the same category.

Interpretaton of results

Sentiment analysis using NLP

We used a pretrained NLP model that showed reliable results. Using this model, we were able to train a linear regression model to predict sentiment based on respondents' ratings. This can be extremely helpful in understanding how loyal consumers are to a brand. Moreover, we conducted a coefficients' impact on sentiment analysis in a linear regression model for each category. It can also be extremely useful for understanding what weaknesses and strengths a brand has and what exactly needs to be developed through advertising.



Centroids and peripheral brands analysis

We were able to successfully identify the centroids for each category and find the brands that stood out the most. We have seen results that indicate that our analysis works very well. For example, we see that in the category "Cars"Audi has destinctive descriptions (thunderstorm, thunder, lightning), which shows that respondents have strong and similar assisiations with Audi and the company can use this knowledge to launch advertisment, showing this assosiations, to make them even stronger and become one and only brand with such assosiations.

In category "Beauty products" we see some compelling results. For example, we see that Old Spice is a peripheral brand. Why? Because it has strong assosiations that nobody else has. "Biceps, shirtless, delicate" are unique associations that Old Spice was able to create. Result of our model shows that their advertisement works perfectly!

In category "Food and dining"Lays Chips is also a periferal brand, which is actually strange at first glance. But thinking more deeply we see that Lays Chips have special assosiations that they projected to costumers, which makes them unique and special in the niche of "Food and dining".

As two possible marketing strategies that companies can use to increase brand awareness, we can highlight orientation towards the centroid and, in a sense, the opposite - the development towards unique associations with high sentimental evaluations. With a successful implementation of the first strategy, companies will be able to build a chain of associations for their brand in such a way that it will gain a broad coverage similar to the most successful, central players in the market. In the second case, companies will be able to focus on the uniqueness of their product and thereby expand their influence on different consumer segments. Assuming that the surveyed population is representative, the regression built in the first part of the study will help determine the emotional coloring of new words that a company may be associated with. This will greatly assist firms in assessing their current position in the market and avoid the need for expensive market research.

There is, of course, the sweet spot in the middle. Our recommendations might be combined to achieve the perfect balance between the uniqueness and the wideness of the assosiations for the brand. This strategy is very applicable for those companies that are new to the market and aren not yet sure what to prefer.

All the results of peripheral analysis you can see in the table below.

Category	**Brands**	**Distinctive descriptions**
Beverages	Corona	tennis, bronzing, bikini
	Maxwell House	clock, bubble, tennis
	Barefoot Wine	bunch, vine, grape
Clothing products	Under Armour	instrument, offspring, baseball
	Dicks Sports	ball, tennis, basketball
	Gucci	bag, jewelry, accessory
Cars	Pontiac	thunderbolt, lightning, thunder
	Audi	thunderstorm, thunder, lightning
	Jeep	hike, adventure, valley
Health products and services	Aleve	teacher, university, classroom
	Walgreens	cup, zoom, baseball
	Band Aid	unify, patriotism, flagpole
Household Products	Purina	dog, fur, pet
	Pledge	half, teeth, citrus
	Swiffer	tooth, laundry, floor
Food and dining	Nature Made	dish, meat, meal
	Dominos Pizza	laughing, performance, musician
	Lays Chips	musician, traffic, couple
Beauty Products	Old Spice	biceps, shirtless, delicate
	AXE	finelooking, delicate, botanical
	Garnier Fructis	bouquet, botanical, delicate
Department Stores	Marshalls	savings, canvas, watercolor
	Macys	merry, sale, row
	Nordstrom	row, skyline, skirt
Home design and Decoration	Frigidaire	frost, frozen, frosty
	Home Depot	cupboard, floor, interior
	Whirlpool	frozen, waterfall, frosty