**Social Impact Analysis (SIA)**

Version 1.0

Updated on Date – 6 Jan 2020

**Process Document**

Contents

[**OBJECTIVE** 2](#_Toc30515960)

[**APPROACH** 3](#_Toc30515961)

[**Execution Guidelines** 3](#_Toc30515962)

[**METHODOLOGY** 7](#_Toc30515963)

[Phase 1:Data Collection 8](#_Toc30515964)

[**Amazon** 8](#_Toc30515965)

[**Facebook** 9](#_Toc30515966)

[**Twitter** 9](#_Toc30515967)

[Phase 2: Extraction of Theme, Sentiment & Emotions 9](#_Toc30515968)

[**Pre-Processing of Text Data: -** 9](#_Toc30515969)

[**Theme Extraction:** 10](#_Toc30515970)

[**Emotion Extraction:** 11](#_Toc30515971)

[**Sentiment Extraction:** 12](#_Toc30515972)

[Phase 3 - Analysis of Themes and Sentiments 12](#_Toc30515973)

[Phase 4 – Presentation of Results 17](#_Toc30515974)

[**FINDINGS & RESULTS** 19](#_Toc30515975)

# **OBJECTIVE**

Social media offers most brands a unique opportunity to engage with their consumers. Understanding key trends, sentiments and themes that are relevant can help a brand strengthen its image. The broad objective of the project social impact analysis is to

* Shed light on the types of social media conversations (themes and experiences) shoppers/consumers have about a brand
* Demonstrate the power of social media buzz to drive category sales
* Understand key sentiments impacting the brand and comparing it with category
* Track brand sentiments on a periodic basis to ascertain key opportunities and threats
* Identify key Themes relevant to consumers and determine brand association /drivers

# **APPROACH**

Natural Language Processing (NLP) is leveraged to gather social intelligence from various platforms. Some of the platforms included in the project are Amazon, Twitter, Facebook and Instagram. Social buzz from these platforms are gathered using advanced web scraping techniques post which NLP is used to derive sentiments, themes and emotions.

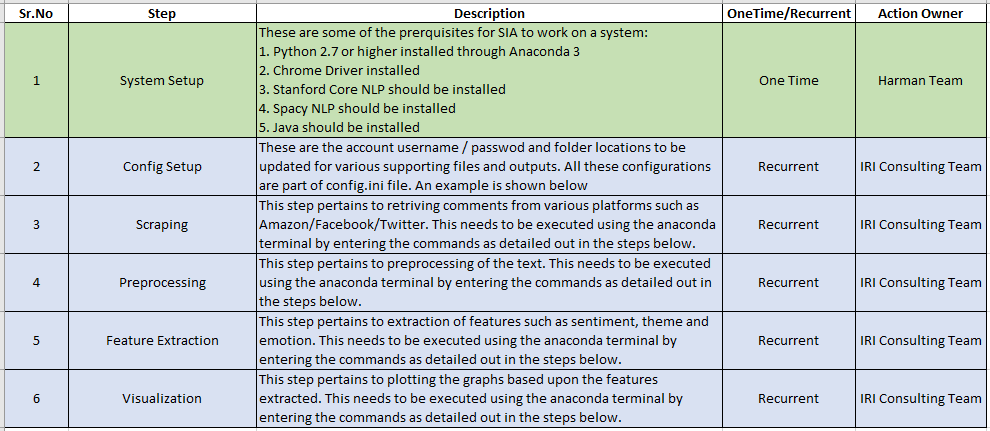
The sentiments, themes and emotions are further analyzed to uncover the hidden patterns using techniques such as correspondence analysis, DNA charts, heat maps and other advances analytics techniques. A mixed effect model is then built to model the relationship of sentiments on sales for each brand and the role played by key themes in the same.

Finally, the findings and conclusions are presented using an interactive dashboard.

## **Execution Guidelines**

This section provides a step wise guideline on how to run the SIA engine. Mentioned below is the high-level process flow followed by the detail of each step:

**Step 1: Python Setup**



System set up is not detailed out here as it will be done by Harman team using remote access.

**Step 2: Config Setup**

[PATHS]

BASEDIR = <base directory where the engine is placed>

CHROMEDRIVER = <sub-directory within base directory where chrome driver is placed>

SUPPORTING\_FILES = <sub-directory with all Supporting files>

[FB\_LOGINS]

USERNAME = <username for fb account for scraping>

PASSWORD = <password for fb account>

**CHROME\_DRIVER**: [Download](https://chromedriver.chromium.org/downloads) the compatible version & place it under

**C:\Program Files (x86)\Google\Chrome\Application\**

**SUPPORTING FILES**: Download from **Whizzer**

**Step 3: Execution**

**Facebook/Twitter:**

It will prompt user to upload a file with a dialog-box to select a file. After file selection, execution will be resumed with the remaining steps

Open command prompt & navigate to the location where all the files are placed.

Then, Execute **main.py**

**Syntax:**

python main.py

Enter the Source: <Platform>

Want to upload the data externally?: <0/1>

Enter the rerun status: <0/1>

Do you want to process Scraping: <0/1>

Do you want to process Preprocessing: <0/1>

Do you want to process Feature Extraction: <0/1>

Do you want to process Clustering: <0/1>

Do you want to process Visualization: <0/1>

Enter the Keyword: <Brand\_Name>

Enter the Number of Pages: <No. of Pages to Scrape>

**Scenario-1(afresh)**: To scrape data from platform & rerun status as No(0)

Enter the Source: twitter

Want to upload the data externally?: 0

Enter the rerun status: 0

Do you want to process Scraping: 1

Do you want to process Preprocessing: 1

Do you want to process Feature Extraction: 1

Do you want to process Clustering: 1

Do you want to process Visualization: 1

Enter the Keyword: starbucks

Enter the Number of Pages: 10

**Scenario-2(upload):** By uploading scraped data externally & rerun status as No(0)

Enter the Source: twitter

Want to upload the data externally?: 1

Enter the rerun status: 0

Do you want to process Preprocessing: 1

Do you want to process Feature Extraction: 1

Do you want to process Clustering: 1

Do you want to process Visualization: 1

Enter the Keyword: starbucks

**Scenario-3(rerun):** By uploading scraped data externally & rerun status as Yes(1)

Enter the Source: twitter

Want to upload the data externally?: 1

Enter the rerun status: 1

Do you want to process Preprocessing: 0

Do you want to process Feature Extraction: 0

Do you want to process Clustering: 1

Do you want to process Visualization: 1

Enter the Keyword: starbucks

**Amazon: Ensure to copy the Review-link file inside <common\_files> folder**

Open command prompt & navigate to the location where all the files are placed.

Then, Execute **main.py**

**Syntax:**

python main.py

Enter the Source: <Platform>

Enter the Keyword: <Brand\_Name>

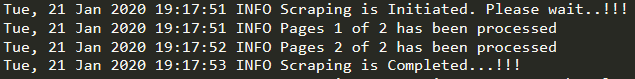
**Example:**

Enter the Source: Amazon

Enter the Keyword: starbucks

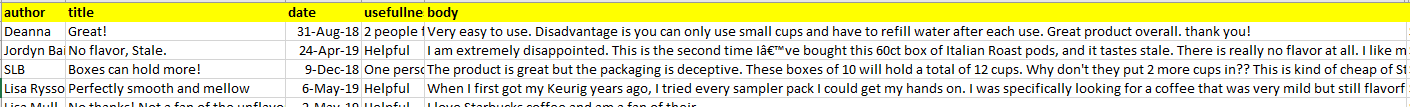
**Step 4: Scraping – Follow the logs**

*Logs:*



*Output:*

A CSV file containing comments along with product name, brand, manufactirer, comment, author name, date, number or likes etc as shown below.



**Step 5: Data Preprocessing – Follow the logs**

*Input: Auto-generated file from Step 4*

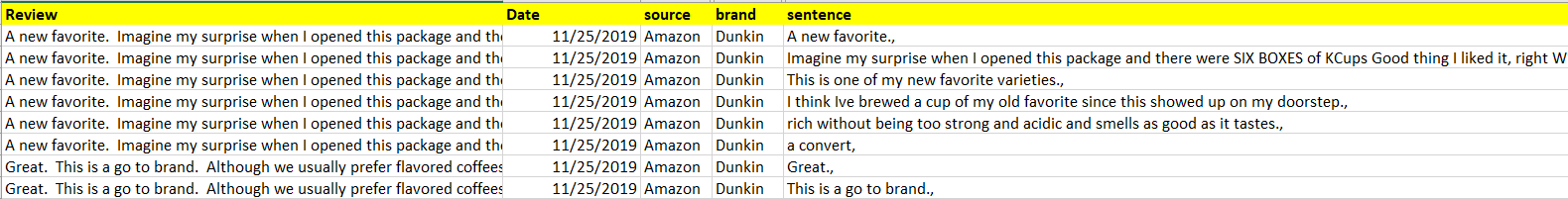
Dataframe containing comments (output of step1 above)

*Logs:*



*Output:*

Dataframe which is cleaned where each comment is broken into sentences. The obtained data will not contain special characters, language other than english, duplicate comments etc.



**Step 6: Feature Extraction – Follow the logs**

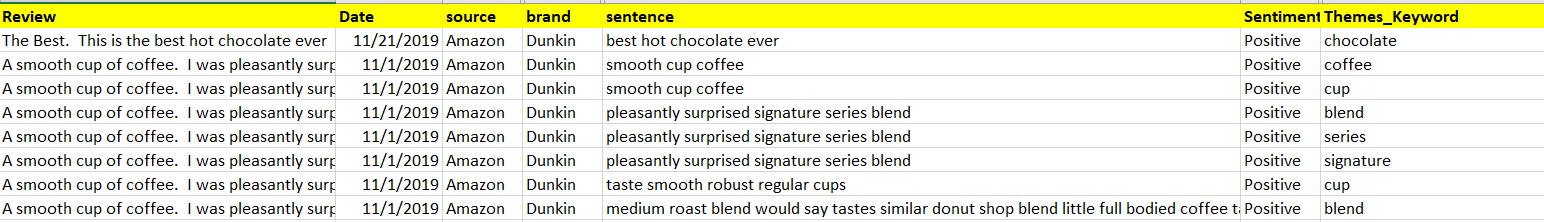
*Input: Auto-generated file from Step 5*

Dataframe containing cleaned sentences (output of step 2 above)

*Output:*

Sentiment (Positive/Negative/Neutral) assigned to each sentence.

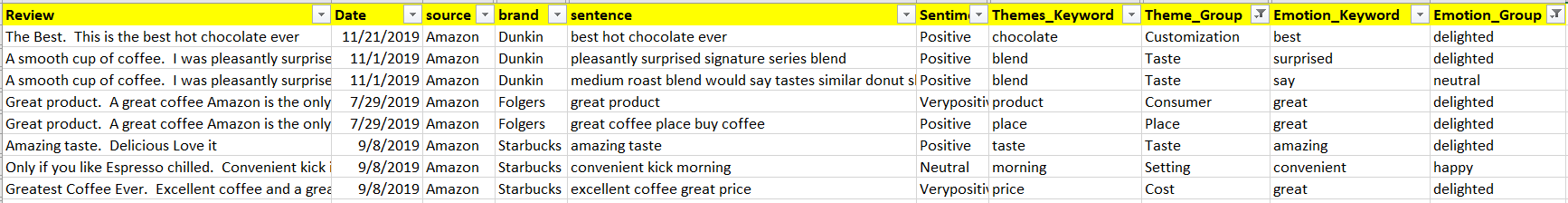
Theme Keyword extracted from each sentence. Please note that there can be more than one theme keyword extracted from a sentence. Also in such case, row pertaining to the sentence will repeat for each theme keyword. An example is shown below:



Emotion keyword extracted for each theme keyword. Please note that emotion is with respect to theme e.g. in statement great taste, bad packaging there would be two theme keywords i.e taste and packaging and there would be two emotion keywords ‘great’ and ‘bad’ associated with taste and packaging respectively.

Finally, all the keywords are clustered into groups. This is done for both: theme and emotion keywords. The mapping of kewwords into cluster can be either provided by user or done in the system itself.

The final output would look like below screenshot:



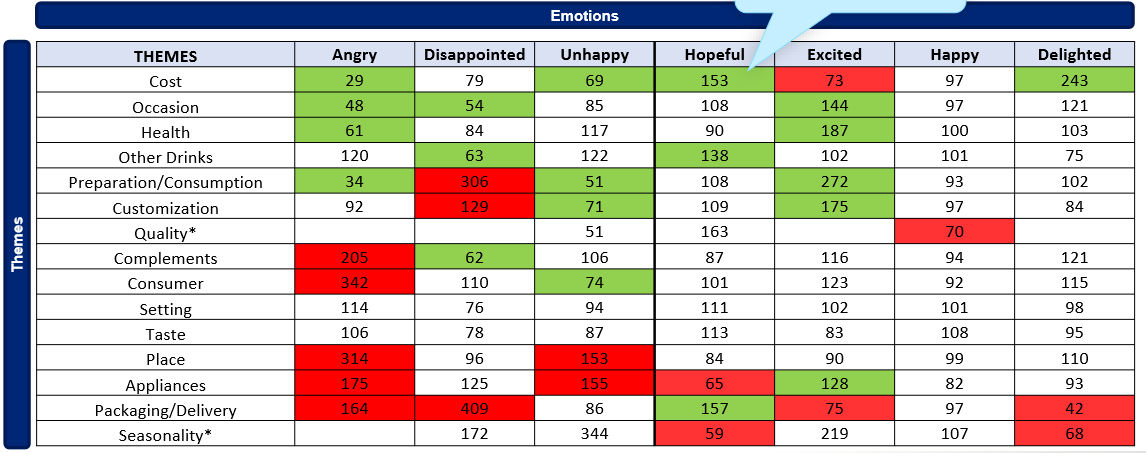
**Step 7: Visualization – Follow the logs**

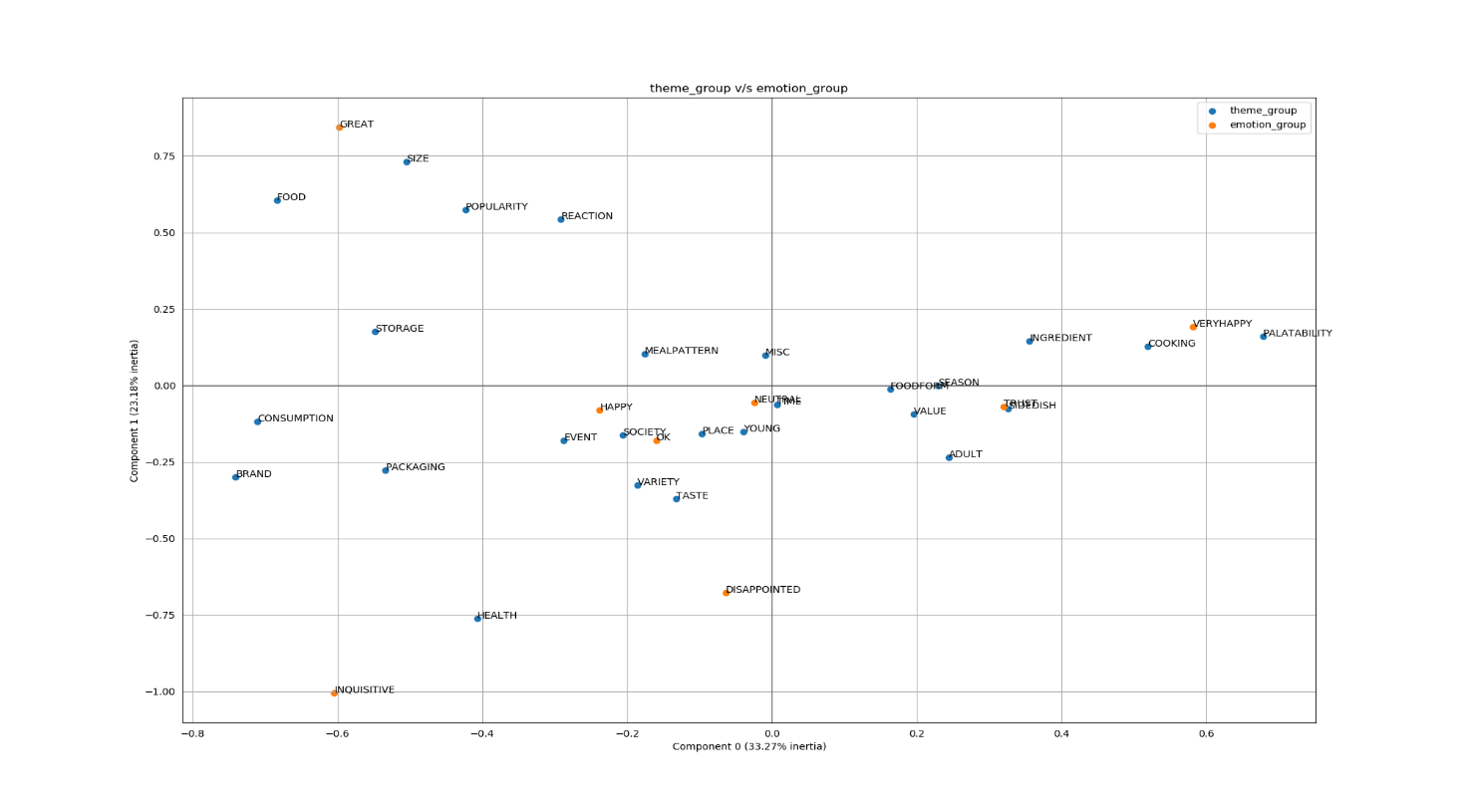
*Input: Auto-generated file from Step 6*

Data frame containing the features extracted in step 3 above.

*Output:*

Various visualization graphs such as word clouds, correspondence chart, heatmaps, line charts etc. A sample is shown below:



# **METHODOLOGY**

**The methodology involves four phases as illustrated below:**

**Phase1: Data Collection –** This phase includes capturing the product reviews and associated demographic information from top online platforms such as Amazon, Twitter, Facebook and Instagram.

**Phase 2: Extraction of Theme, Sentiment & Emotions –** Phase 2 involves various NLP tasks such as tokenization, stemming, lemmatization etc., followed by extraction of sentiments. After deriving sentiments, theme keywords are extracted and are further used to determine the associated emotion keywords. Both theme keywords and emotion keywords are then clustered using word embeddings.

**Phase 3: Analysis and Visualization –** In phase 3, the sentiments and themes derived from phase 2 are analyzed and findings are generated. Statistical techniques such as correspondence analysis, DNA charts, regression models and mixed effect models are used to draw meaningful, actionable and generalizable insights.

**Phase 4: Presentation of Results –** The insights generated from phase 3 are presented using interactive dashboard and the same is published to the user.

## **Phase 1:**Data Collection

In this phase mainly we focus on collecting the data from top online platforms such as Amazon, Facebook and Twitter to get the product reviews and related demographic information.

Amazon

Twitter

Facebook

**Data Collection**

Scrapping

**Amazon**

We use python codes to get reviews or review comments by customers from Amazon. The approach here is targeted to focus on the specific brand /product page to extract the reviews. Complete process of scrapping data from amazon is automated.

We have automated the process of data scrapping from amazon using python scripts and Libraries like selenium and beautifulsoup.

**Details of getting or scrapping data from amazon is explained in below steps:**

* Go to Amazon Home page
* Search using keyword - Product name for e.g. Cadbury Dairy Milk Silk
* Go to the particular product using UPC and scrap below information:
  + Name of Product
  + Product Manufacturer
  + Rating on Product
  + Price of Product
  + All the review along with below details:
    - Title
    - Body
    - Author Name
    - Star rating given
    - Date of review
    - How many found it useful
    - If purchase verified (Y/N)
    - Demographics and associated details like reviewer rating, #friends, #followers etc.

### **Facebook**

We have a semi-automated process to scrap data from Facebook. Here we use python codes to scrap data and store them in csv files.

**Process of scrapping data from Facebook is explained below:**

* Provide the keyword to the Facebook scrapper function e.g. “”
* Use the post URL obtained from step one above to scrap the following information:
  + Comments written on the post
  + View and like on comment
  + Name, location and number of friends of the author

### **Twitter**

To scrap the data from twitter we follow process similar to Facebook.

## Phase 2: Extraction of Theme, Sentiment & Emotions

### **Pre-Processing of Text Data: -**

**The steps involved in pre-processing of the text data are shown in the below figure:**

**Cleaning - The first step in pre-processing of the reviews captured in phase 1 is data cleaning. We remove special characters, numbers, punctuations and other irrelevant symbols and characters such as tabs, empty spaces, new line characters etc. We also remove the common stop words. Some of the common stop words in English are “This”, “The”, “That”, “is” etc.**

**Tokenization - The review comments are then tokenized into sentences. We also create a sequential ID for each review so that a sentence can be traced back to a review. Tokenization is important because as per our analysis we have found that a review comment is generally made up on many sentences and there are multiple themes and sentiments in a review, pertaining to each sentence. An example of tokenization is shown below:**

**"Because I could not stop for Death, He kindly stopped for me. The Carriage held but just Ourselves and Immortality"**

**Sentence 1 - Because I could not stop for Death, He kindly stopped for me.**

**Sentence 2 - The Carriage held but just Ourselves and Immortality"**

**Stemming - Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of reducing it to common base form. As an example of stemming the word “Played”, “Playing” and “Play” will all be replaced by the common base form “Play”.**

**Lemmatization - Lemmatization also works like stemming however, it usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. As an example of lemmatization, words “is”, “are” are both replaced by “be”.**

**Named Entity Recognition – Using pretrained models, named entity i.e. personal nouns are identified in the sentence and removed.**

**We use Stanford Core NLP, NLTK and Spacy to perform the above tasks.**

### **Theme Extraction:**

**Stanford Core NLP is used to derive themes. The steps involved in theme extraction are mentioned below:**

**Part of Speech Tagging using Stanford Core NLP – Tokenized sentences are parsed using Stanford Core NLP to tag part of speech.**

**Rule Based Engine to Derive Theme Keywords – Post tagging the part of speech a rule-based engine is used to derive theme keywords. This engine has been validated using a huge dataset and the performance is evaluated in terms of recall and coverage.**

**Recall of the engine – Recall is defined as the percentage of keywords which are tagged by the engine as theme keywords out of relevant theme keywords.**

**Recall = (Theme Keywords Derived ꓵ Relevant Theme Keywords)/Relevant Theme Keywords**

**Coverage of the engine – Coverage is defined as the percentage of sentences for which the engine has derived at least one theme keyword.**

**Coverage = Number of Sentences assigned a Theme Keyword / Total Number of Sentences**

**K Means Clustering of Theme Keywords using Word Embedding -** After tagging theme keyword(s) to each sentence, k-means clustering is done using word embedding and each theme keyword is grouped into a cluster. Glove embedding are used for clustering. These embedding are pretrained on huge corpus of data as a result of which the cosine similarity between similar words high and thus k-means clustering tends to place them in similar groups. An example is shown below:

50 Dimensional Embeddings

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Texture** | **1.23** | **1.09** | **0.9** | **0.76** | **-0.33** | **….** |
| **Mushy** | **-0.98** | **1.23** | **1.67** | **1.34** | **-0.99** | **….** |
| **Soft** | **1.11** | **1.43** | **0.88** | **0.33** | **1.4** | **….** |
| **Hard** | **0.13** | **0.65** | **-0.3** | **-0.89** | **2.5** | **….** |
| **Value** | **1.67** | **1.11** | **1.43** | **-0.98** | **1.23** | **….** |
| **Cost** | **0.88** | **0.00** | **-0.67** | **1.11** | **1.43** | **….** |
| **Cheap** | **-0.3** | **1.11** | **1.43** | **0.13** | **0.65** | **….** |

**Naming of Cluster using wordnet and Domain Expertise – After clustering, the clusters are named using wordnet and the same is validated by domain experts.**

WordNet

+

Domain Expertise

….

….

### **Emotion Extraction:**

The emotion keyword is extracted based upon theme keyword. For each keyword, the relationship with different words in the sentence is analyzed and the words expressing attitude/emotion of the theme are extracted. For example, in the sentence – “We love the taste”, the word “love” expresses the emotion with regard to the theme “taste”. After extracting the emotion keywords, they are clustered in the manner similar to the one described above (**K Means Clustering of Theme Keywords using Word Embedding). The result of clustering is a set of emotion groups assigned to each theme.**

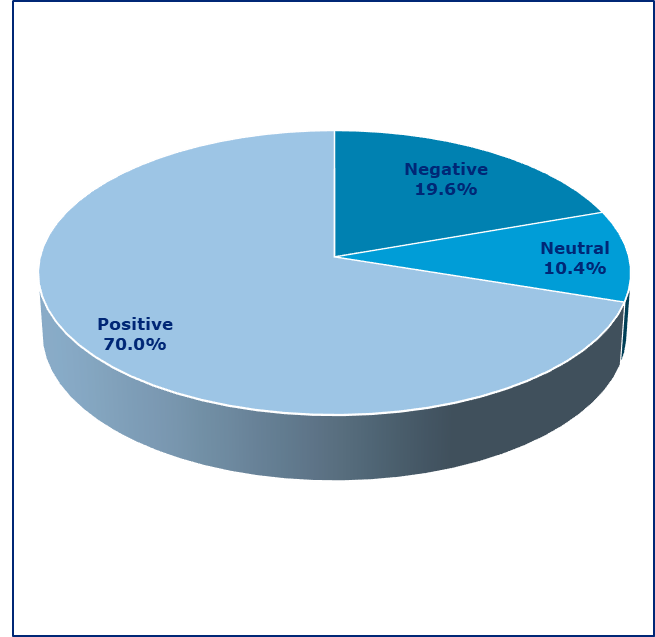
### **Sentiment Extraction:**

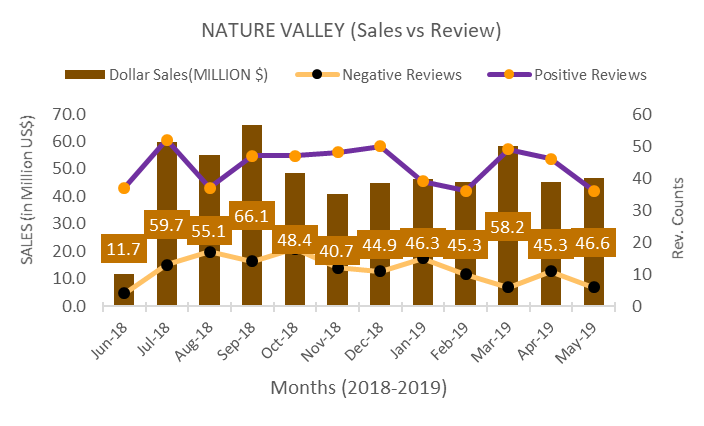
In order to derive sentiment of a sentence, pretrained deep learning models are used. A combination of Stanford Core NLP and textblob is currently used to tag sentiment. The accuracy is evaluated on regular basis to make modifications.

## Phase 3 - Analysis of Themes and Sentiments

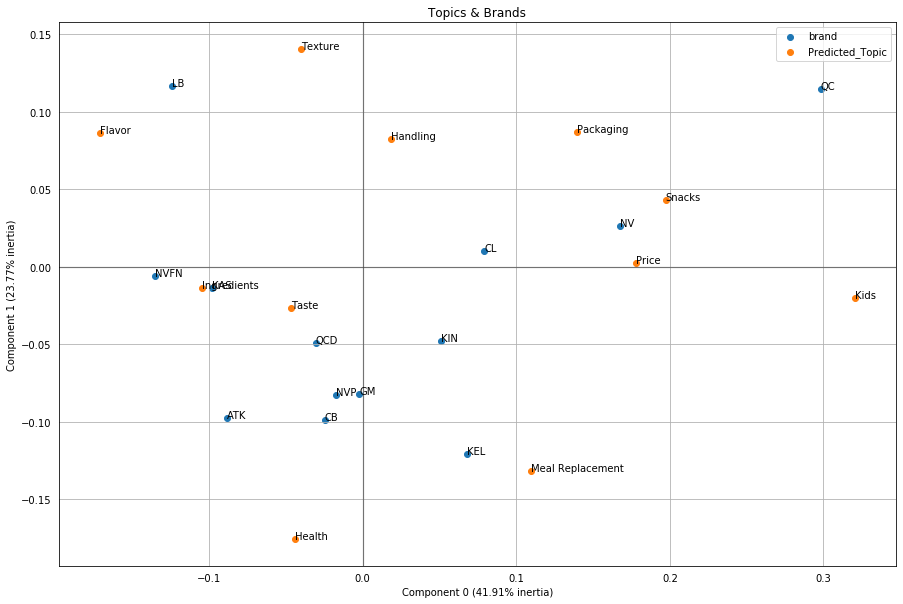
**The polarity and themes obtained using deep learning models are then analyzed using the below mentioned techniques to uncover the hidden patterns.**

1. **Exploratory Analysis** – The purpose of exploratory analysis is to uncover the patterns such as polarity associated with brand and theme, distribution of comments across brands, distribution of comments across themes, trend associated with a theme and words making up a theme etc. The exploratory analysis is done using various visualization techniques as shown below;

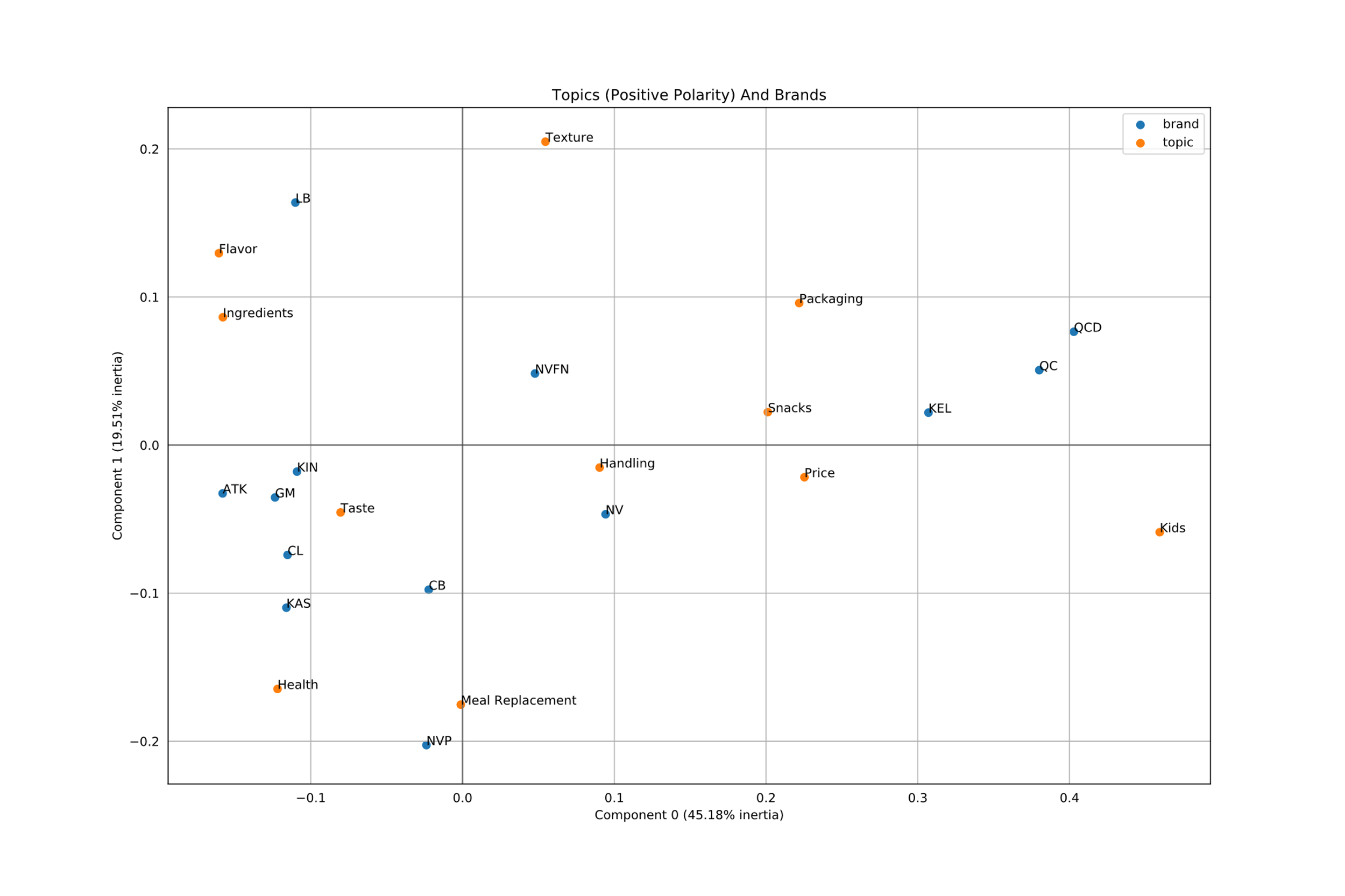
1. **Correspondence Analysis: - We use Correspondence analysis when we want to study the dependency between two or more categorical variables. Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by Factor Analysis techniques, and they allow you to explore the structure of categorical variables included in the table. In phase 3 we study the relation between “Brand” and “Themes” using correspondence analysis. . We used correspondence analysis in conjunction with the heat map to understand how different brands are associated with various themes. An example of correspondence analysis and heat map is shown below:**



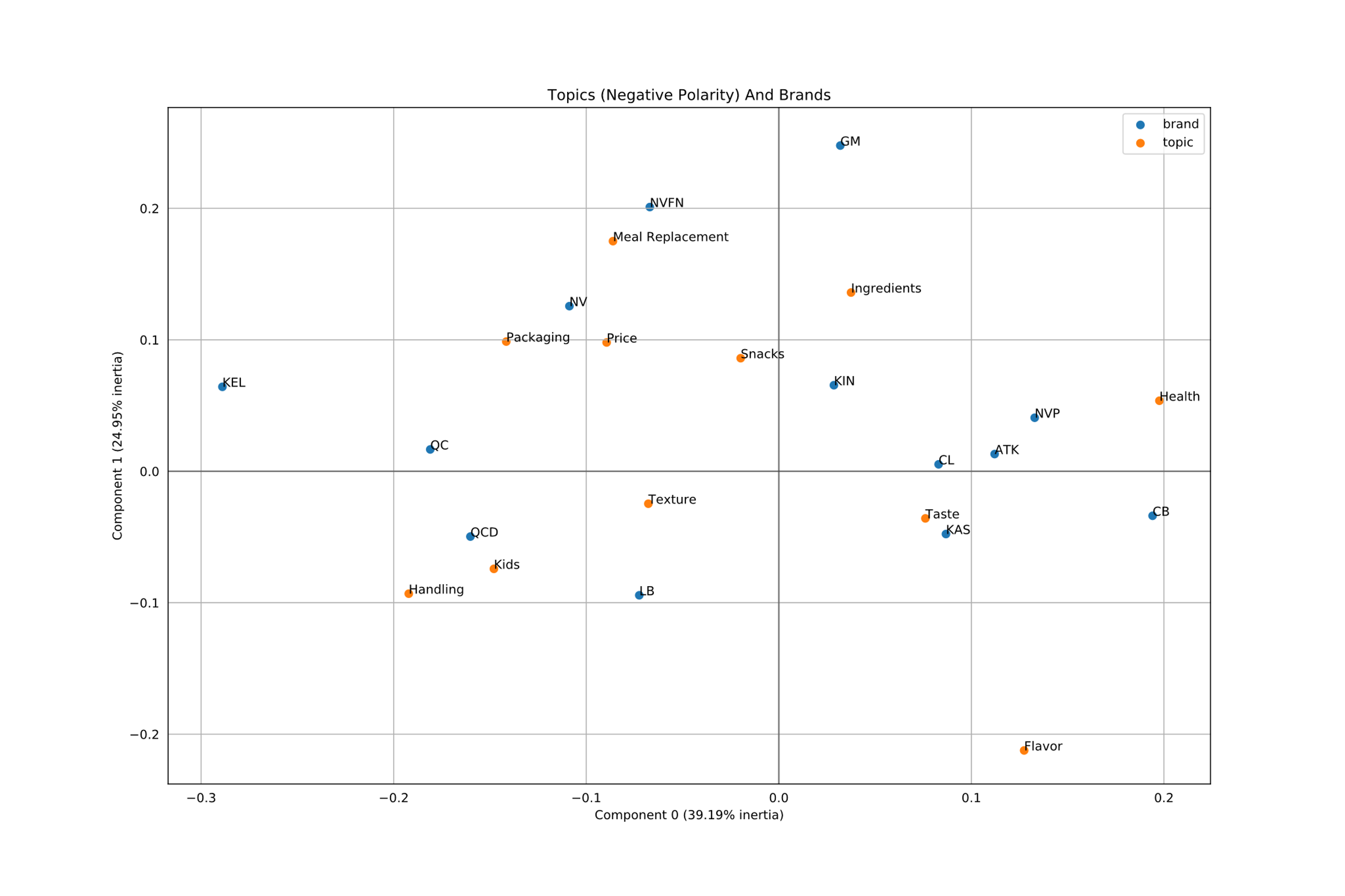
**Further we also use correspondence analysis to understand how different brands are related to different themes based on polarity associated with theme. In order to achieve this, we split the overall comments based upon polarity and then use correspondence analysis to see how different brands relate to various themes. This helps us understand why people are happy or unhappy with a brand.**

**Shown below is an example of correspondence analysis done separately for positive and negative sentiments.**

**Correspondence Analysis for Positive Sentiments**

**

**Correspondence Analysis for Negative Sentiments**



**If we look at the two plots, we can see that the in both cases the two dimensions explain a variance of about 64%.**

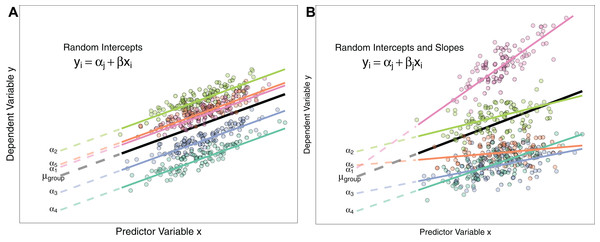
**We cross-verified the perceived relation-ship (looking at the above plots) between “Brand” and “Product Attributes” by constructing a residual table. A residual value shows the associations between the row and column labels. Big positive numbers mean a strong positive relationship. The opposite is true for negatives.**

1. **Regression Analysis: - In order to understand the impact of sentiments on sales, we regress the sales on Net Positive Sentiments (NPS). NPS is derived by subtracting the count of negative sentiments from positive sentiments for a period. An example of regression modelling is shown below:**

**Using the ordinary least square we found that** each positive sentiment is associated with increase in 0.08 Million US Dollar Sales on a monthly basis.

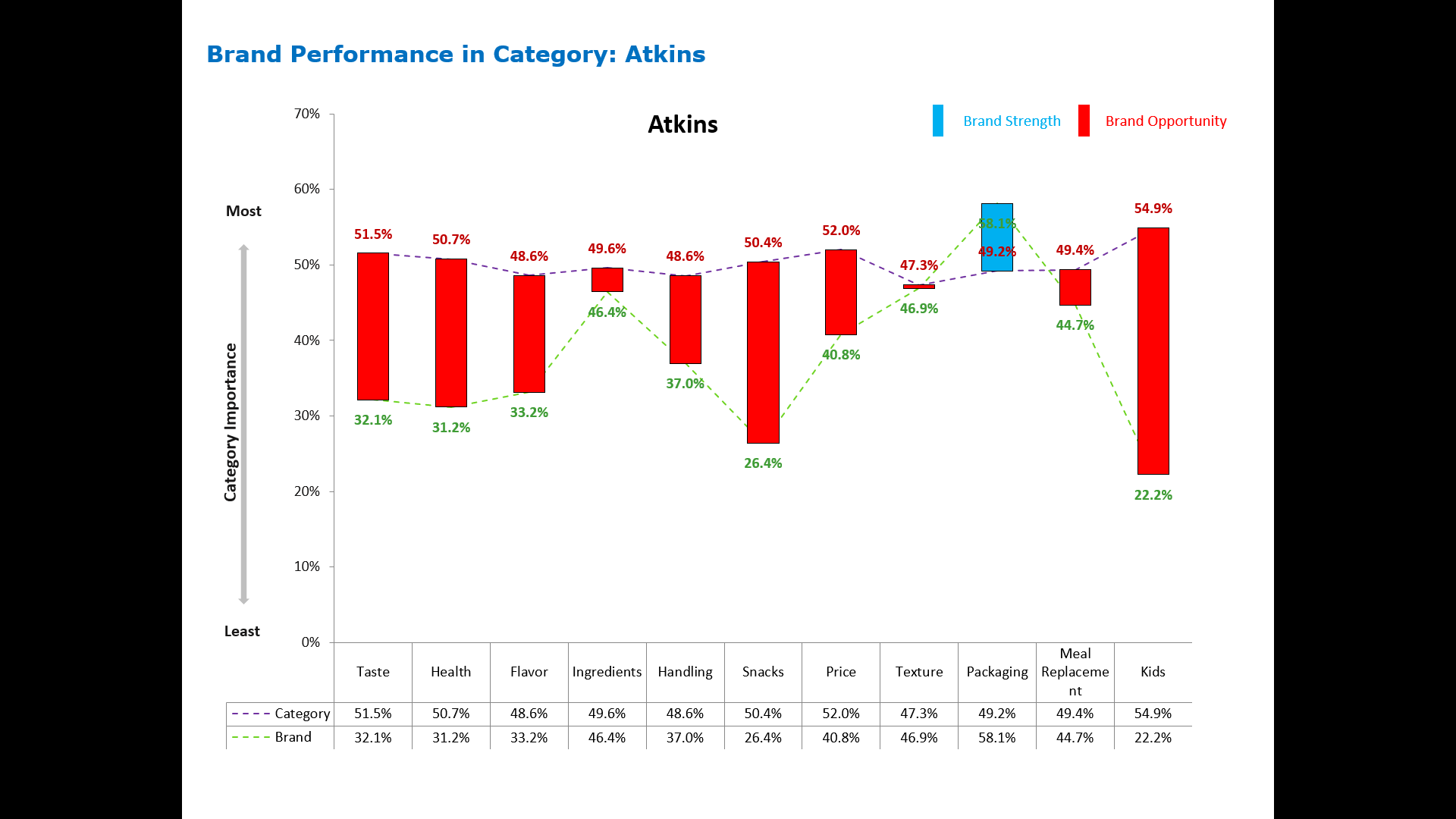


1. **Mixed Effect Model** - We further analyze the relationship of NPS with each of the brand along with themes through mixed effect model. This is used on panel data (time series + cross sectional data) and can help prioritize the improvement efforts by providing monetary value associated with each theme.



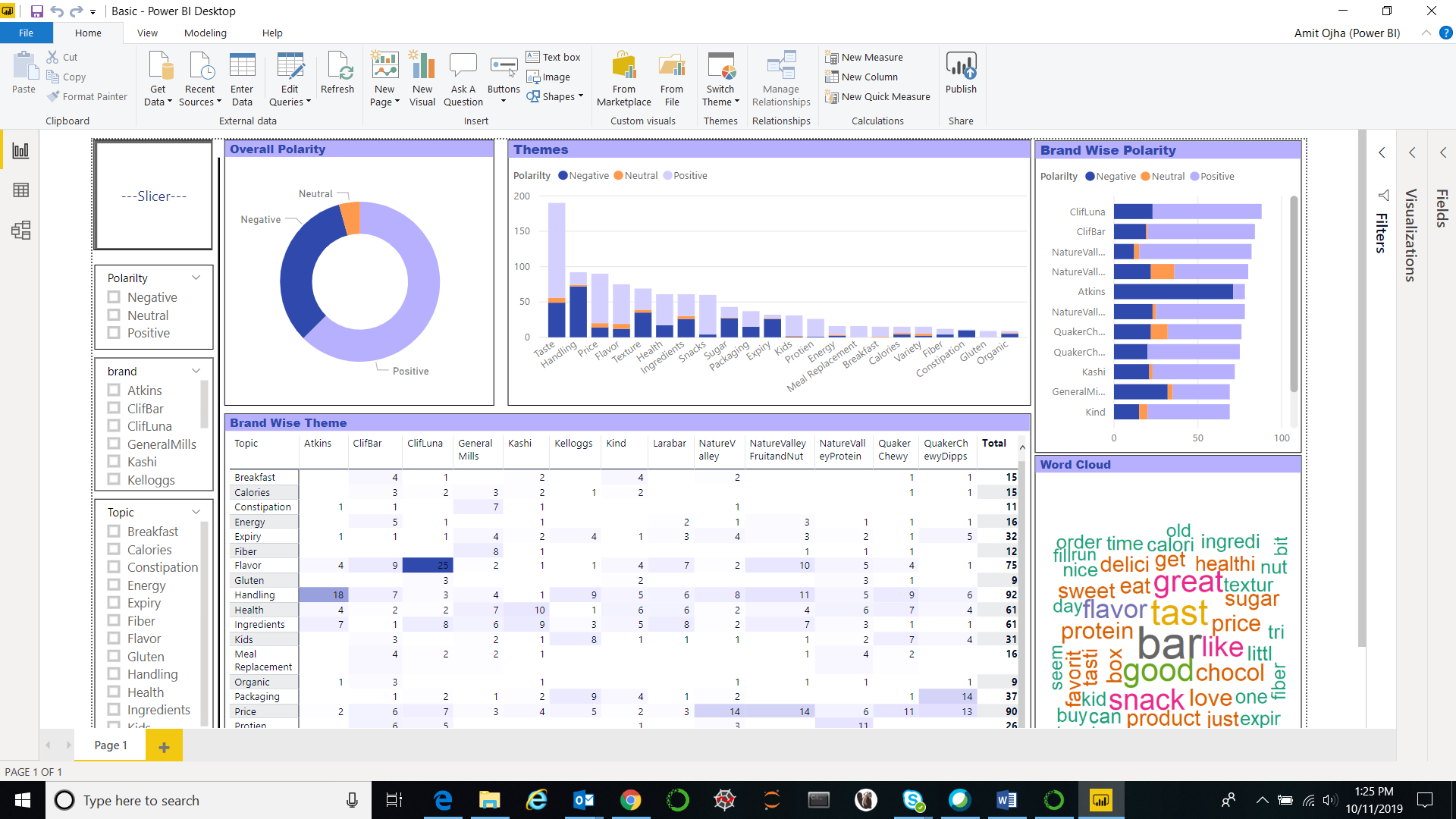
(<https://peerj.com/articles/4794/>)

1. **DNA Charts** – The purpose of DNA charts is to compare the performance of two or more classes with respect to predefined category. We use DNA charts to compare the performance of each brand with category on various themes. These charts are easy to interpret & can convey a lot of information in a concise form. An example of DNA chart is shown below:



## Phase 4 – Presentation of Results

The insights obtained in phase 3 are integrated into an interactive dashboard using Power BI. An example is shown below:



# **FINDINGS & RESULTS**

**Mentioned below are some of the results from this project which can help a brand strengthen its image:**

* Identify the emotions associated with overall category and various brands.
* Discover the key consumer preferences for brands and classify them in logical groups. As an example, some of the preferences associated with snack and granola bars are mentioned below:
  + **Feel** - Flavor, Texture
  + **Brand Perception** – Price, Value, Packaging
  + **Benefits** – Meal Replacement, Ingredients
  + **Demographics** – Kids, Adults
* Evaluate the positioning of a brand by understanding the relationship of a theme with customer attitude
* Find the relationship between monthly sales trend and net positive sentiments
* Identify the key attributes for each brand to focus for growth opportunities