2021

Building a Brand Perception Model with Twitter Data

ISYE6740 – TEAM NODJO

PRASHANT KUBSAD (903608055), KRISHNA KUMAR GANESH (903616390) & NEVI SHAH (903130252)

TABLE OF CONTENTS

INTRODUCTION:

In this new age of social media, a whole new meaning has been brought to effective consumer marketing and research. Gone are the days when marketing consisted only of designing singular ads, commercials and emails because today it is fueled by intense market research on current and potential consumers, driving more personalized experiences and products with great market fit. Companies today utilize a breadth of technologies in order to appeal to and understand the market of their company. With more than 3.96 billion people utilizing social media, analyzing social networking data is a fast and easy way for companies to get free and both structured and unstructured data about their consumers. Some of the most important insights drawn by this sort of analysis includes a comprehensive understanding of their audiences such as to who they are, what their interests are, where they come from and how they react to certain trends, products and ideas. Social media has been a consistent source of real time data giving accurate depictions of target consumers and allows companies to predict the adoption of new products and services.

Twitter in particular has been an instrumental contribution to every company's understanding of their consumer's attitudes. Consumers are always quick to turn to Twitter to discuss dissatisfaction of their experiences, specifically as it relates to areas like customer preference, in-store experience and online experience. Not only does this help companies to understand and document these experiences when it is negative but also enables them to collect positive anecdotal evidence with respect to the brand. This naturally informative way of sharing customer insight is the crux of business development and market research and it is completely free to companies if the company has a means to extract the data and build its own models. Because of the rise in popularity of Twitter data analysis, Twitter has provided a series of APIs allowing companies to pull Tweet data as a raw dataset enabling them to see information like geographic location, Tweet text, number of followers and more. Company data scientists are then able to extract this data but are then required to cleanse, normalize and build a model for the data on their own. A big challenge that companies face is trying to find a quick and automated way to digest and derive actionable insights on live Twitter data without having to read each tweet and manually document it. In this paper, we will investigate the Coca-Cola Company (Coca-Cola) and all tweets with the #cocacola hashtag. At a high level, we will extract the Twitter data feed for this hashtag and develop an algorithm that utilizes both sentiment analysis and clustering to understand Coca-Cola's customers' biggest pain points as well as areas of strong positive opinion.

BACKGROUND:

Coca Cola is a total beverage company based in Atlanta, Georgia with its products sold and purchased in 200+ countries across the globe. While known for its "Coca Cola Classic" drink, the Fortune 500 company also carries a plethora of sports drinks, water, coffees, teas, and other concentrates/syrups. Founded in 1892, Coca Cola has gained widespread popularity for its canned sodas that are localized in taste depending on the geographic location of business. However, in order to cater to the needs of every region of customers, Coca-Cola invests a significant amount of money in performing consumer market research.

One of the most famous Coca-Cola consumer research cases dates back to the 1970s in the "Race Against Pepsi". Though Coca-Cola had a significant lead in the market share, Pepsi released a campaign revealing consumers preferred Pepsi every time in a blind taste test because it was sweeter. As a result of this study, Coca-Cola released "New Coke" a sweeter version of regular Coke, however, it was one of the biggest failures causing a revolt among customers. The reason it failed was because in a blind taste test, the cola isn't being paired with the meals consumers normally drink it with and only a few sips were given

instead of a whole can. Coca-Cola learned that while user studies can be helpful, they do not always paint the full picture.

Fast forward 40 years, the company has learned its lesson to keep constantly monitoring the tastes and preferences of its consumers in order to address and iterate on its current products. Today, consumers are quick to turn to Twitter in order to describe their affinity toward or distaste for the products on the shelves. Live Twitter is the perfect way to understand what consumers are saying, thinking and feeling about Coca-Cola products at real time without spending millions of dollars on focus groups and studies.

PROBLEM STATEMENT:

How might we derive the areas of Coca Cola consumer dissatisfaction by utilizing Twitter data?

DATASET OVERVIEW:

In order to address this problem statement, we will be requiring two data sets: (1) Twitter sentiment training data and (2) Coca Cola specific Twitter data.

TWITTER SENTIMENT TRAINING DATA

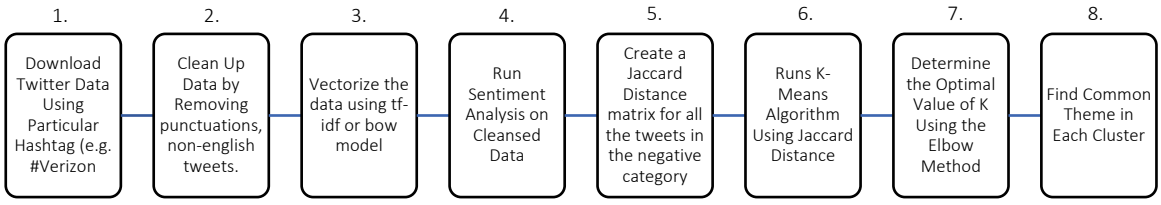
In building our model, developing training and testing data to build the model was one of the biggest challenges faced. We decided on using an already labeled dataset made available by Plural Sight to train/test our initial model for sentiment analysis. This dataset consists of over 400,000 tweets not specific to any company or user and has already been conveniently labeled with a binary value classifying each tweet in the data set as either positive or negative.

COCA-COLA SPECIFIC TWITTER DATA:

After we have trained our model using this prelabeled dataset, we also had to pull the Coca-Cola tweets to answer our decided problem statement. Because we wanted to gather all tweets addressing Coca-Cola, only tweets with the hashtag #cocacola were pulled to avoid any other noise in the dataset. This dataset was pulled on April 28 2021, and contains 2,080 rows of tweet data. This data is the crux of our sentiment and cluster analysis and can be replaced by any company's dataset of tweets.

METHODOLOGY:

In approaching this problem, the following workflow has been established:



CLEANSING

After downloading our datasets, we first started by cleansing and normalizing our data because we cannot just pass in raw data. To begin the cleansing process, we followed a series of procedures to prepare the tweet for vectorization. Because tweets are unstructured data, it is crucial for us to be able to create consistency in each tweet and make sure we are valuing the right words and pieces of the tweet for our sentiment analysis. For example, taking the example tweet shown to the right, below is the cleansing process performed:
I love the new cherry flavor!

Cleansing Step	Output
Remove URLs	I love the new cherry flavor! The taste is so incredible and refreshing! @CocaCola #cocacola #softdrinks
Remove Hash Symbol	I love the new cherry flavor! The taste is so incredible and refreshing! @CocaCola cocacola softdrinks
Remove @ Mentions	I love the new cherry flavor! The taste is so incredible and refreshing! cocacola softdrinks
Convert to Lowercase	I love the new cherry flavor! the taste is so incredible and refreshing! cocacola softdrinks
Remove Punctuations	I love the new cherry flavor the taste is so incredible and refreshing cocacola softdrinks
Remove Stop-words (library provided by NLTK.corpus)	love new cherry flavor taste incredible refreshing cocacola softdrinks

VECTORIZATION

Next, the data must then be vectorized, which converts words and word tokens into numbered vectors. This step is crucial in quantifying English words in order to run any machine learning algorithm on them. To perform vectorization, we were found two different effective approaches: (1) Bag of Words model and (2) Term Frequency -Inverse Document Frequency (TF-IDF).

BAG OF WORDS MODEL

This is the simplest form of converting text representation to numbers and extract features from text. It describes occurrence of words within a document. In order to build the bag of words we first developed a vocabulary of known words. This step involves constructing a *document corpus* which consists of all the unique words in the whole of the text present in the data provided. It is often compared to something like a dictionary where each index will correspond to one word and each word is a different dimension.
For example, below are two tweets pertaining to Coca-Cola (note these examples do not reflect cleansed tweets such that stop-words/punctuation/etc. are removed):

Tweet #1

Tweet #2

If we collect the unique words from the above two tweets and assign each to an index, we will have:

Document Corpus

Index	1	2	3	4	5	6	7	8	9
Word	This	cherry	flavor	is	very	tasty	and	affordable	not

After we have constructed a document corpus, we will then record a count of the presence of our known words. For example, if we take Tweet #1 and plot the count of each word, we will have a row that corresponds to the index of the unique words and a row that corresponds to the count.

Word Count: Tweet 1

Index	1	2	3	4	5	6	7	8	9
Count	1	1	1	1	1	1	1	1	0

After converting both tweets into such vectors we can compare different sentences and calculate the Euclidean distance between them to check if two tweets are similar or not. Thus, if there are no words in common the Euclidean distance would be much larger and vice-versa.

Term frequency-inverse document frequency model (TF-IDF)

The next approach we found to vectorizing our data is called TF-IDF with two distinct parts (1) Term Frequency and (2) Inverse Document Frequency.

Term Frequency (TF) is used in connection with information retrieval and shows how frequently a word occurs in a tweet. Term frequency indicates the significance of a particular term within the overall document.

$$tf_{t,d} = \frac{n_{t,d}}{N}$$

$n_{t,d}$	$number\ of\ times\ the\ term\ 't'\ appears\ in\ document\ 'd'$
N	$number\ of\ terms\ in\ document.$

Inverse Document Frequency (IDF): The inverse document frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score. It is the logarithmically scaled inverse fraction of the documents that contain the word.

$$idf_{t,D} = log\ \frac{N}{|\{d \epsilon D\ :\ t \epsilon d\}|}$$

$\|d \epsilon D\ :\ t \epsilon d\|$	$number\ of\ documents\ where\ the\ term\ 't'\ appears$

Thus, TF-IDF is calculated as:

$$tf - idf_{(t,d,D)} = tf_{(t,d)} . idf_{(t,D)}$$

As shown from the equation, TF-IDF gives larger values for less frequent words in the document corpus and vice versa. TF-IDF value is high when both IDF and TF values are high i.e the word is rare in the whole tweet dataset but frequent in a particular tweet.

SENTIMENT ANALYSIS WITH CLASSIFICATION

As mentioned, our approach was to use a predefined dataset that has more than 400K tweets that are already marked with 0 or 1 (0 being negative sentiment, 1 being positive sentiment). After vectorizing these 400K tweets using both above methods of vectorization, we split the dataset into train (80%) and test (20%) data. We trained the data for classification using two different models: (1) Multinomial Naïve Bayes and (2) Logistic Regression.

MULTINOMIAL NAIVE BAYES CLASSIFIER:

Naive Bayes classifier is best suited for text classification, spam filtering and sentiment analysis. Naive Bayes classifier works on the Bayes rule, given by:

$$posterior\ probability = \frac{conditional\ probability * prior\ probability}{predictor\ prior\ probability}$$

$$P(\frac{A}{B}) = (\frac{P(A \cap B)}{P(B)}) = \frac{P(A) * P(\frac{B}{A})}{P(B)}$$

$P(A)$	$the\ prior\ probability\ of\ occuring\ A$
$P(B/A)$	$the\ condition\ probability\ of\ B\ given\ that\ A\ occurs$
$P(A/B)$	$the\ conditional\ probability\ of\ A\ given\ that\ B\ occurs$
$P(B)$	$the\ probability\ of\ occuring\ B$

Multinomial Naïve Bayes implement the Naive Bayes algorithm for multinomially distributed data. The distribution is parametrized by vectors $\theta y = (\theta y1, \dots, \theta yn)$. The parameters θy is estimated by a smoothed version of maximum likelihood, i.e., relative frequency counting:

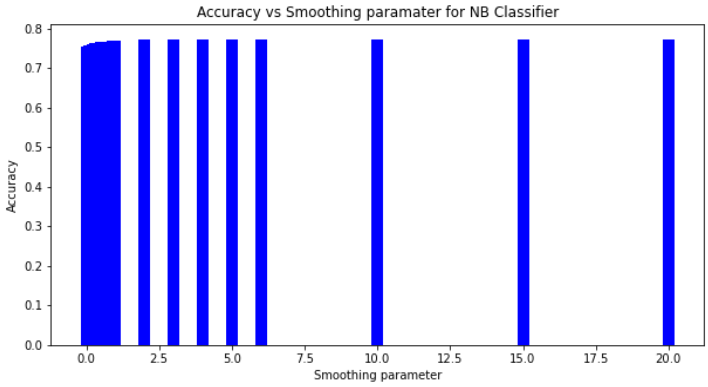
$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

$\theta y = (\theta y1, \dots, \theta yn)$	each class y, where n is the number of features (in text classification, the size of the vocabulary)
θ_{yi}	the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y.
$N_{yi} = \Sigma_{x \epsilon T} x_i$	is the number of times feature i appears in a sample of class y in the training set T
$N_y = \sum_{i=1}^{n} N_{yi}$	total count of all features for class y
α	$\alpha \geq 0$ accounts for features that are not present in the learning samples and prevents zero probabilities in further computations.

We have used the Sk-learn package's MultinomialNB method to run the above logic with different values for smoothing parameters. We opted to go with a smoothing parameter of 15 because it becomes consistent after $\alpha = 2$ as shown in the graph below.

MultinomialNB with Different Smoothing Parameters

```
alpha:  0.0 accuracy:  0.75344375
alpha:  0.1 accuracy:  0.758434375
alpha:  0.2 accuracy:  0.76111875
alpha:  0.3 accuracy:  0.762990625
alpha:  0.4 accuracy:  0.7644875
alpha:  0.5 accuracy:  0.765525
alpha:  0.6 accuracy:  0.76621875
alpha:  0.7 accuracy:  0.7668875
alpha:  0.8 accuracy:  0.7675125
alpha:  0.9 accuracy:  0.768034375
alpha:  1.0 accuracy:  0.768521875
alpha:  2 accuracy:  0.771090625
alpha:  3 accuracy:  0.77195625
alpha:  4 accuracy:  0.772075
alpha:  5 accuracy:  0.7723625
alpha:  6 accuracy:  0.772471875
alpha:  10 accuracy:  0.772459375
alpha:  15 accuracy:  0.772553125
```



LOGISTIC REGRESSION

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. As learnt in the classes, logistic regression classifies data using probability of a data belonging to a category. We define a threshold for this probability of data, and if it crosses the threshold, we can safely classify it to that category. Logistic regression follows sigmoid function:

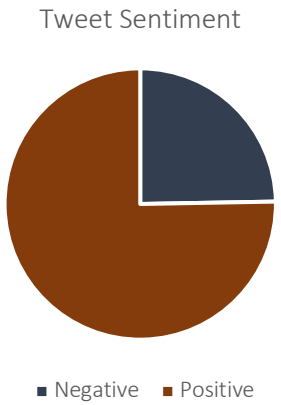
$$y = \frac{1}{1 + e^{-x}}$$

SENTIMENT ANALYSIS RESULTS

Combining all the approaches we tried the following combinations and recorded the accuracy of each:

Vectorization	Classification	Accuracy
TF-IDF	Multinomial NB	0.772553125
TF-IDF	Logistic Regression, max -iter 250	0.787696875
TF-IDF	Logistic Regression, max -iter 500	0.788209375
Bag of words	Multinomial NB	0.773078125
Bag of words	Logistic Regression	0.78565

As shown, all the models with their respective combinations returned very similar accuracy rates. Looking at these five, we chose the model that yielded the highest accuracy even if just by a small margin. This combination was TF-IDF vectorization and Logistic Regression classification with a *max-iter* of 500. Next, we apply this model to our Coca-Cola Twitter data which has been cleansed and vectorized per the methods above, we can see that the model classifies tweets into 369 *negative* Tweets and 1124 *positive* tweets. See below for examples of positive and negative tweets.



Negative Tweets	Positive Tweets
RT @glbl301 : These are only a few instances, but there are many, many more. The more I looked up, the more cases there seemed to be of #Nestlé and other beverage giants like #CocaCola and #Pepsi depriving water poor communities around the world of access to clean drinking water.	4PM ET - The New Retail master class and networking w/ #CocaCola Barry Thomas Microsoft @ricardo_belmar #Salesforce @MGTalksRetail Proximis @KhamtanThavy @cathymccabe Proximity Insights @Winston_W_Ma global tech investor and more @rwthurston1 Retail Pride
"RT @PatrioticSocia1 : It's okay to be white.There is nothing wrong with it.#cocacola#BoycottCocacola"	Beverage giant joins a growing list of global corporations sponsoring a protected reserve in the Amazon. #FMTNews #CocaCola
RT @fordmb1 : The right amount of #CocaCola is zero. Not Coke Zero, but Zero Coke. #Woke #WokeBreakingPoint #boycottcoke #BoycottCocacola #Hypocrites https://t.co/7lme0lTsmv	RT @OriginalFunko : This Pop! is sure to be the cherry on top of your collection. Coming Soon: Pop! Funko: @CocaCola - Cherry Coca-Cola can🍒. Pre-order yours now, Cheers!

CLUSTER ANALYSIS

Clustering falls under the realm of unsupervised machine learning algorithms – We do not know before-hand the different clusters that are present in the data, and we expect the algorithm to give us insights into the different clusters available. While it is not straightforward to identify the best way to cluster data, there are multiple algorithms readily available to assist with this – We chose the K-means algorithm as our clustering algorithm to help perform the clustering of similar tweets. As we learned, the efficiency of the K-means clustering is sensitive to the initial starting points for each of the cluster centroids – it is easy to start with some bad cluster centers and have clustering results which are not easily decipherable. Thus, we tried two different ways to assign the initial cluster centroids for the K-means algorithm and compared the outcomes from both: (1) Having K-Means automatically select the best cluster centers randomly and (2) Use the Jaccard Distance metric to identify the tweets which are extremely uncorrelated with each other to identify the initial center points.

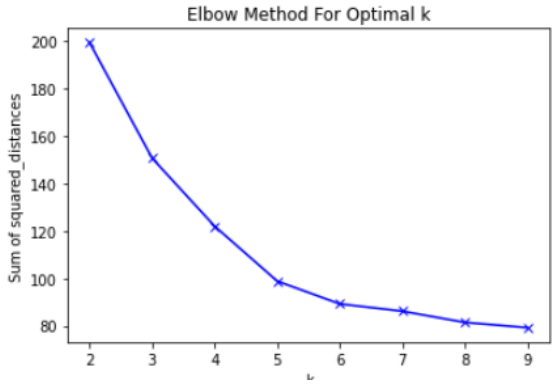
DATA CLEANSING AND PREPARATION

The data for clustering analysis is the negative tweets identified by sentiment analysis step above. We run similar data cleansing steps as above resulting in a vector matrix like:

Word	TF-IDF Score
water	0.304960
many	0.269237
are	0.269237
more	0.260634
Glbl301	0.188627
access	0.176928

FINDING THE K-VALUE

We initially tried multiple K-values manually to identify the proper number of clusters that would make the clustering meaningful using 'k-means++' initialization of cluster centroids. We found that K = 2, 3 and 4 were too low of values. Clusters were very unevenly sized, with majority of tweets falling into just one cluster and only a few tweets falling into other clusters. While K = 10, 11, and 12 were too high, as we could not decipher any meaning from these high number of clusters since there were too few tweets falling into each cluster. For a more accurate K-Value we finally utilized the elbow plot method in order to observe the infamous "kink in the curve". Based on the plot to the right it looks as if K= 5 or K = 6 seems like the appropriate number of clusters. In trying both, we can see that K=5 is a more evenly distributed cluster. Below is a table of the distribution of each cluster at each of the K values:



Cluster Number	Number of Tweets Per Cluster (K=5)	Number of Tweets Per Cluster (K=6)
0	276	276
1	61	93
2	36	61
3	30	36
4	104	30
5	-	11

CENTROID INITIALIZATION

In choosing the method of centroid initialization for these clusters, there were three main possible routes to take:

1. **Choosing 'k-means++' as the initialization parameter of SK-Learn K-means to assign centroid points:** As per SK-learn documentation, the 'k-means++' centroid initialization selects initial cluster centers for k-mean clustering in a smart way to speed up convergence. We tried this method to see if the K-means algorithm out of the box produces a good clustering or not.

2. **Choosing initial centroid points using the Jaccard Distance calculation of centroid points:** The Jaccard index is also called the Jaccard similarity coefficient is used in understanding the similarities between sample sets. While the Jaccard Index calculates the SIMILARITY between two sets the Jaccard Distance measures the DISSMILIARITY between two sets and is defined as Jaccard Distance = (1 - Jaccard Index). Formulaically, it can be written as:

$$Jaccard\ Distance = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

   In this method, we calculated the Jaccard Distance between pairs of tweets from the entire tweets collection to identify tweets which were totally dissimilar from each other. As a rule of thumb, pairs of tweets with the least Jaccard Distance estimate proved to be reasonable selection for our initial cluster centers. Once the cluster centers were identified, our K-means algorithm performed the below:

1. Initialize cluster centroids to the 5 most dissimilar tweets using Jaccard's distance
2. For each incoming document (i.e tweet vector), calculate the cosine similarity between the tweet and the tweet at each of the cluster center.
3. Assign the tweet to the cluster which has the most similarity with the cluster centroid
4. Repeat this process until all the tweets have been processed.

3. Choosing 'random' as the initialization parameter of SKLearn Kmeans to assign centroid points: As per SKLearn documentation, the 'random' centroid initialization selects K rows at random from data for the initial centroids. We also tried this method to identify if the clustering gets better if the centroids were chosen totally at random.

EVALUATION/FINAL RESULTS

The output of the K-means clusters was fed into a Word Cloud package, which helped us visualize the different key words that were prominent across the different clusters. Based on our experiments above, we observed that the overall clustering performed by the below 3 methods were similar – but the K means algorithm where we identified initial clusters using Jaccard Distance metric did a better job of clustering more tweets into each cluster, thus giving the word cloud better interpretability. Interestingly, the K means algorithm where centroids were chosen at random did a similar clustering as the K-means algorithm with 'k-means++' method of centroid initialization.

Below are how the cluster sizes looked like for the negative tweets of Coca Cola after each of these algorithms. Please note that we cannot compare the cluster counts 1:1 between each of the methods due to randomness incorporated in the K-means heuristic.

	K-Means++ centroid	Jaccard	Random
Cluster1	276	35	104
Cluster2	61	67	36
Cluster3	36	10	276
Cluster4	30	34	61
Cluster5	104	366	30

CLUSTER RESULTS AND ANALYSIS

We looked at the different tweets which were on each of the clusters from the K-means algorithm to identify what the defining themes were within each of the clusters. Below were the 5 different themes we were able to capture out of the clustering from the negative tweets for Coca Cola. We have added an overview of the problem/pain point as well as potential strategies/solutions Coca-Cola could use in response to these issues. This can be considered as a way for Coca-Cola to take this analysis and produce actionable insights from them. Following this table are the topic results for each cluster along with the associated Word Cloud.

Cluster	Topic of Cluster	Analysis
1	Coca-Cola's racially insensitive training	• **Problem**: Coca Cola consumers care about racial inequality and care deeply about social justice • **Solution**: Release a statement clarifying or addressing this issue to avoid losing loyal customers.
2	Coca-Cola stock becoming more expensive	• **Problem**: This is actually good news for Coca-Cola as a company but not great for its consumers who are looking to invest • **Solution**: Release a new wave of marketing initiatives helping hungry investors make the decision and justify investing in Coca-Cola stock.
3	High price of Coca-Cola products	• **Problem**: Perhaps increasing the prices of Coca-Cola products may be increasing sales at the moment but may not be sustainable if this is how consumers think • **Solution**: Perform a pricing strategy audit involving sales/finance/operations to understand how much Coca-Cola can lower the price without losing a lot of margin.
4	Reaction to Coca-Cola marketing campaigns	• **Problem:** This marketing campaign heavily stresses that consumers are not a fan of Coke Zero. • **Outcome:** This may give Coca-Cola the starting point to investigate if they should be switching Coke-Zero out for a new product on the shelves.
5	Coca-Cola localized taste used as metaphor to democracy by political leader	• **About**: There is no inherent pain point in this tweet as Coca-Cola was mentioned for a metaphor • **Outcome**: However, the metaphor for Coca-Cola flavor and democracy is a concept that seemed to resonate with a lot of consumers and might be something Coca-Cola could use for future marketing campaigns given the positive success of the tweet.

CLUSTER 1: BAD PRESS DUE TO RACIALLY INSENSITIVE TRAINING

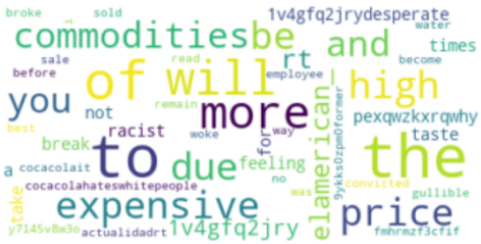
Analysis: The most common tweet/retweet within this cluster was *"RT @PatrioticSocia1 : It's okay to be white.There is nothing wrong with it.#cocacola#BoycottCocacola"*. It looks like this retweet/negative sentiment was broadly due to a training conducted by Coca Cola where they suggested employees be "less white"

(https://www.entrepreneur.com/article/366132).

CLUSTER 2: COCA COLA STOCK

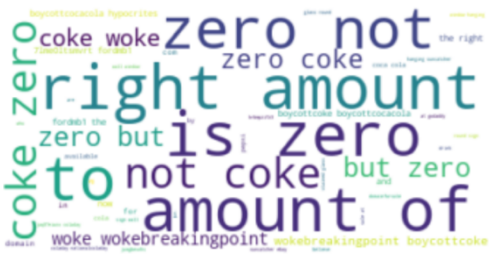
The tweets within this cluster were all along the lines of investing and suggestions about selling Coca Cola's stock. The tweets seem to suggest that even Coca Cola's stock price is going to be more expensive, and buying crypto currencies is the cheapest option today. One of the sample tweets from this cluster is *"RT @CryptoJoker12 : don't be trapped in the short-term volatility. just a year ago #Bitcoin was $5k and #xrp at $0.17. Inflation is coming. Even #CocaCola gonna be more expensive. In such a market, I don't think #Gold #Silver #BTC #DGB #xrp #xlm #cryptocurrency gonna be any cheaper."*

CLUSTER 3: EXPENSIVE PRODUCTS



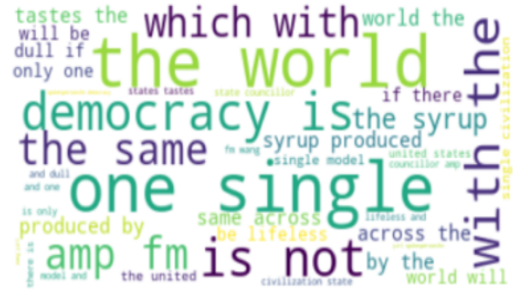
Most of the tweets seem to have fallen within this cluster and so this cluster shows a variety of patterns. However, most of the negative sentiments within this cluster seems to be on the high price of some of the coke products with key words such as expensive, price, high and commodities.

CLUSTER 4: MARKETING CAMPAIGN REACTIONS



This one was an interesting cluster as most of the tweets within this cluster were in response to a marketing campaign by Coke where they asked folks on twitter what the right amount of ice in Coca Cola is. The Tweet for reference is: RT @fordmb1: The right amount of #CocaCola is zero. Not Coke Zero, but Zero Coke. #Woke #WokeBreakingPoint #boycottcoke #BoycottCocacola #Hypocrites https://t.co/7lme0lTsmv

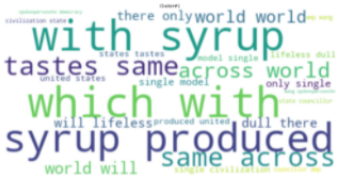
CLUSTER 5: POLITICAL COMPARISON BY FOREIGN MINISITER WANG YI

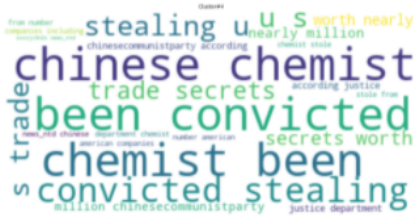


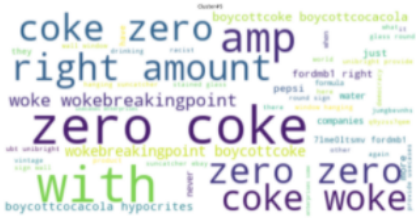
The tweets within this cluster were mostly centered around a comment made by Chinese Foreign Minister Wang Yi, which was retweeted multiple times. The tweet for reference is: RT @SpokespersonCHN: #Democracy is not #CocaCola, which, with the syrup produced by the United States, tastes the same across the world. The world will be lifeless and dull, if there is only one single model and one single civilization. --State Councillor & FM Wang Yi

CLUSTERING USING K-MEAN++ METHOD

For comparison, below are the clusters we got for K-Means ++ method as well as the random initialization method (we found them to be exactly the same). However, the clusters above used Jaccard Distance. We can see a similar trend in the clusters for both Jaccard and K-Means++ in terms of topic yet our team found the Jaccard clusters to be more meaningful/clear regarding category of pain point. To repeat this analysis our recommendation would be to execute using the Jaccard distance.







CONCLUSION/LESSONS LEARNED

After performing both the sentiment and cluster analysis on this Coca-Cola dataset, we can see that this model provides extremely valuable insights into the consumers and their preferences. Each cluster has given us a key takeaway from its consumers, whether that be that Coca-Cola drinkers care deeply about race and equality, or that marketing campaign are an extremely successful option to create chatter on or that perhaps the Coca-Cola products are too expensive for consumers and the pricing strategy needs to be reevaluated. Of course, Coca-Cola could send out user research surveys and coordinate focus groups, but all of that takes time and money. This algorithm gives Coca-Cola and other consumer-oriented companies like it the ability to get insights fast and their fingertips. In addition, this sort of quick analysis allows companies to not only understand their consumers better but quickly take control of the narrative. For example, regarding the racially insensitive training that was given by Coca-Cola, this analysis allows the company to fully understand who are tweeting these tweets, what they are saying, and in doing so, curate an appropriate response for each of the tweets causing the most defamation.

While our goal was to make this model something that was reproducible and applicable to all different types of company Twitter data, one of the biggest lessons learned on this project is that not all hashtags are unique and telling of a certain company. If this model is intended for understanding brand perception it is important to understand that hashtags may not be the best identifier for selecting data. For example, if Apple Inc. wanted to perform this analysis for their brand, #apple may not be the best way to filter through Twitter information as the word apple can also just be a regular fruit bought at the supermarket. For this reason, our recommendation going forward would be perhaps to identify tweets that are directed at the company using an @apple mention and to tailor tweets even more, additionally filter on tweets that have a certain hashtag such as #iphone12. This way, data scientists of the model can eliminate and noise in the data regarding tweets that have nothing to do with the brand.

DIVISION OF WORK

This team evenly split all the work to pull this project together. Each team member was present for each part of the decision-making process from ideation to development to report generation and completion. We are proud of the work we have put forth together and had an enjoyable experience with this project overall.

SOURCES/REFERENCES:

https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes
https://www.pluralsight.com/guides/building-a-twitter-sentiment-analysis-in-python
https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22
https://machinelearningmastery.com/gentle-introduction-bag-words-model/
https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/
https://kambria.io/blog/logistic-regression-for-machine-learning/#:~:text=Logistic%20regression%20is%20a%20classification,either%20a%200%20or%201
https://ieeexplore.ieee.org/document/5340335
https://towardsdatascience.com/clustering-documents-with-python-97314ad6a78d

SOURCE CODE

https://github.com/Prazhant/ISYE6740_Project