

# NLP Analysis of Amazon Question and Answer banks

CAPSTONE PROJECT

BOB SPOONMORE

OCT 2021



[HTTPS://WWW.BIZJOURNALS.COM/COLUMBUS/NEWS/2017/08/03  
/REVIEWING-AMAZON-RESTAURANTS.HTML](https://www.bizjournals.com/columbus/news/2017/08/03/reviewing-amazon-restaurants.html)

# Problem Identification:

## Do the Questions Asked by Customers in the Amazon Food Category Receive Correct Answers?

An NLP modeling comparison between different methods of text identification on unlabeled data to determine if questions were answered



[HTTPS://WWW.AMAZON.COM/FMC/LEARN-MORE?REF\\_=PRIMENOW](https://www.amazon.com/fmc/learn-more?ref_=primenow)

A Capstone project:

- Amazon Question and Answer banks
- Unlabeled Text Fields
- Data: 2018 Data Extraction for Amazon category: Grocery and Fine Food
- Predicting the probability of user question being answered correctly

# Cost of Not Answering Customer Questions

3

Churn Rate: (Attrition Rate) The percentage of customers that stop using a service

In 2020, Statistica reported that for Online Retail the churn rate was 22%,  
twice the rate of similar, big box retail stores

In 2020, Statistica reported that total online sales for Amazon in online food  
and beverage in the United States as \$20.37 Billion

The Potential loss rate for Amazon not answering customer questions in the  
category of food and beverage in the United States is \$4.5 Billion

A leading cause for customer dissatisfaction is unanswered questions

Statistica, 'Customer Churn Rate by Industry', <https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/>

Statistica, 'Amazon-us-grocer-gmv', <https://www.statista.com/statistics/545906/amazon-us-grocery-gmv/>

# Data

Source Dataset: [https://jmcauley.ucsd.edu/data/amazon/qa/qa\\_Grocery\\_and\\_Gourmet\\_Food.json.gz](https://jmcauley.ucsd.edu/data/amazon/qa/qa_Grocery_and_Gourmet_Food.json.gz):

Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems

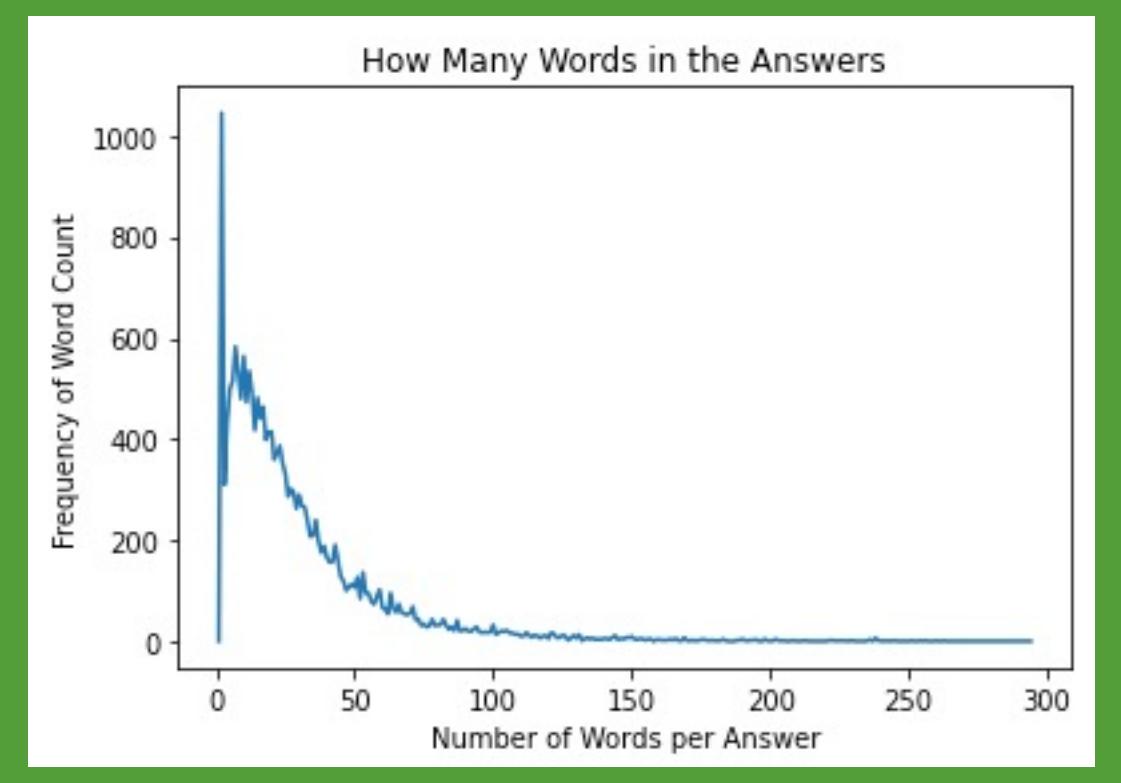
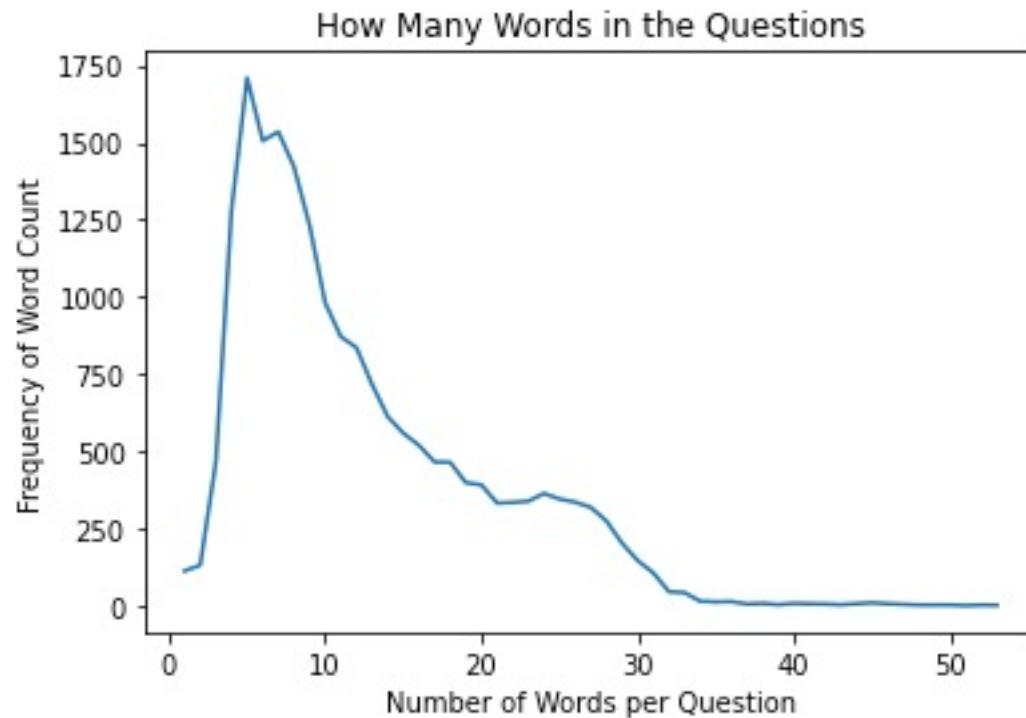
Mengting Wan, Julian McAuley International Conference on Data Mining (ICDM), 2016

Addressing complex and subjective product-related queries with customer reviews Julian McAuley, Alex Yang World Wide Web (WWW), 2016

	questionType	asin	answerTime	unixTime	question	answer	answerType
0	open-ended	9742356831	Mar 26, 2014	1.395817e+09	What is the heat of this compared to the yellow? Is there MSG in it?	I think that the yellow is the most mild. The ... No MSG in Mae Ploy curry pastes.	NaN N
1	yes/no	9742356831	Apr 2, 2014	1.396422e+09	what are the ingredients exactly in this product?	The ingredients are listed in the description!	NaN
2	open-ended	9742356831	Apr 5, 2015	1.428217e+09	How important is the expiration date on the product?	I never pay attention to it myself. The ingred...	NaN
3	open-ended	9742356831	Aug 19, 2014	1.408432e+09	The product description says 14 oz., but the package	We bought the 14oz for just under \$5.	NaN
4	open-ended	9742356831	Aug 2, 2014	1.406963e+09	How many pieces are in a bag?	It depends on the size of the pieces.	NaN

shape: (19538, 7)

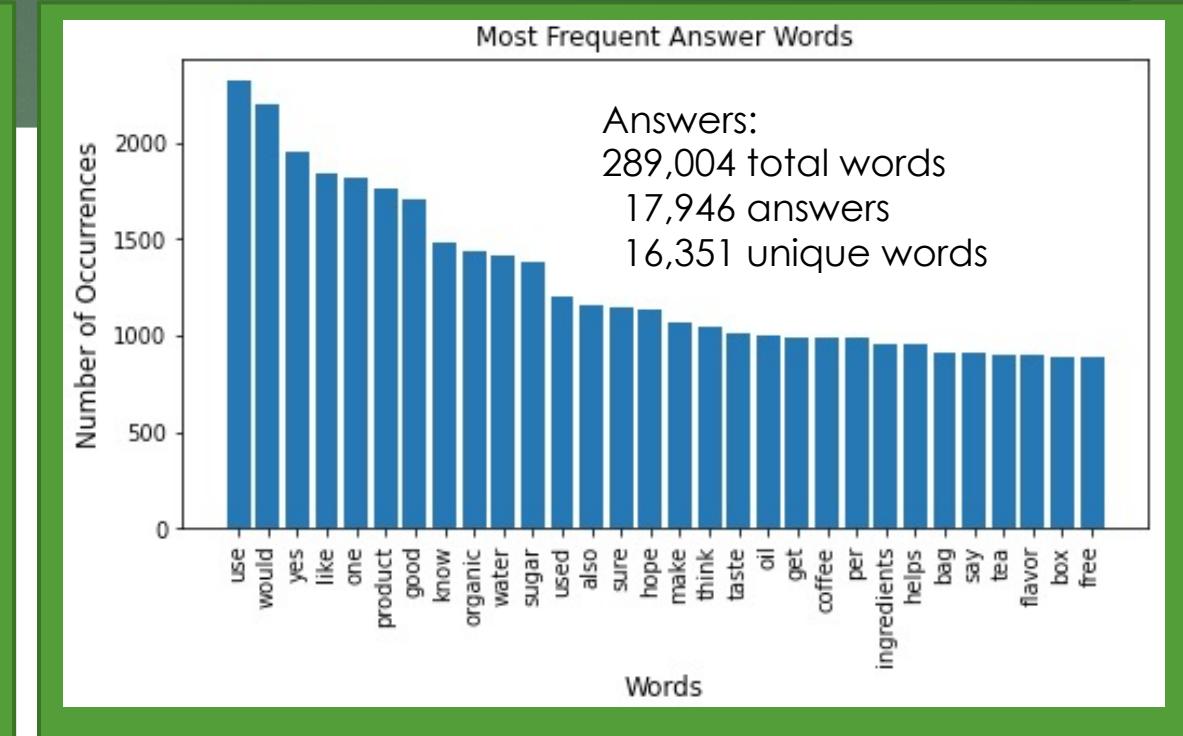
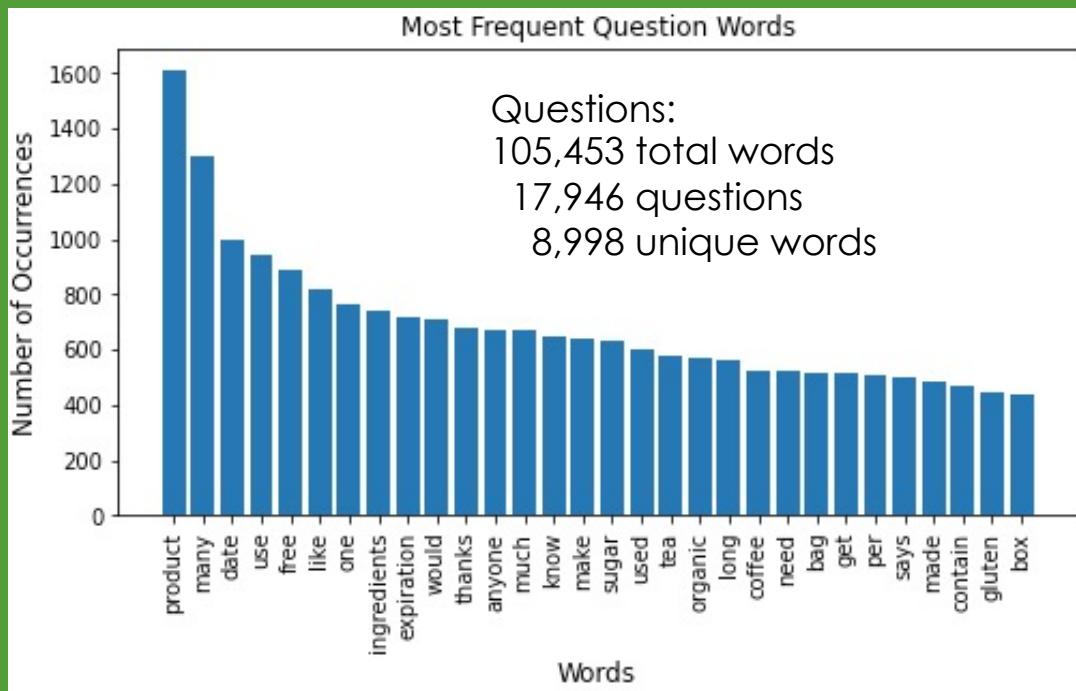
# Exploratory Data Analysis



No null values found

Removed all 1 or 2 word questions and answers as these would bias models

# Question Words versus Answer Words



Frequency distribution of words found in questions and answers

# Preparing Text for Model

## Text Prep on Both Questions and Answers

- Preconditioned text: set lower case, remove punctuation, and remove stopwords
- Kept Questions and Answers separated, and had total group combined
- Created Bag of Words Dictionary and Corpus based on combined list
- TF-IDF matrix to downgrade most frequent words
- Created Bigrams listing

### Highest Frequency Question Bigrams

- Hamilton\_beach
- Dolce\_gusto
- Trader\_joe
- Agave\_nectar
- Genetically\_modified

### Highest Frequency Answer Bigrams

- Chocolaty\_refer
- Possibilities\_jimmies
- Aspergillus\_oryzae
- Drift\_pollinators
- Sri\_lanka

# TF-IDF Analysis (Sklearn)

Term Frequency - Inverse Document Frequency takes bag of words corpus and down weights words that appear most frequently

## Modeling

- 1) Bag of words from Q & A
- 2) Fit and Transform to vectorize
- 3) Combine Corpus, Dictionary to TF-IDF weights
- 4) List Top Values

### Questions

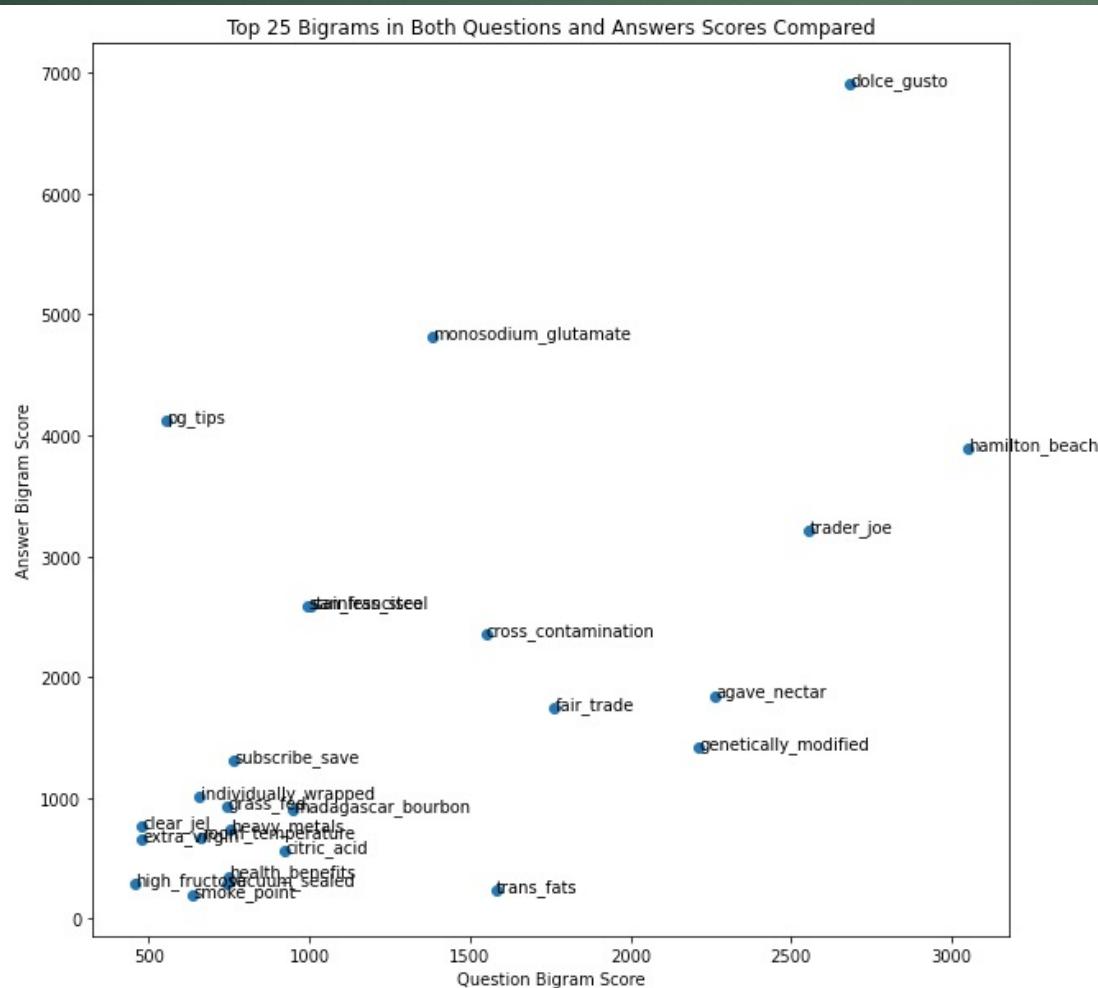
**Top feature Names:**  
curry 0.5280719034995475  
red 0.37460475940536125  
yellow 0.4292527979030419  
compared 0.4501747390819909  
heat 0.4403363245862047

### Answers

**Top feature Names:**  
red 0.24963019415811116  
profile 0.40116616287683743  
flavor 0.2004650839356598  
deeper 0.42718542776135887  
green 0.2503323670052263  
mild 0.3294613696529366  
yellow 0.5893069023938735  
think 0.18863489360450084

# Bigrams of Text

Common Terms Associated with each other – combined as one



The bigrams shown have some clustering with similarity between Q and A,

However, some outliers show a bias

### Higher Answer Use:

- Dolce\_gusto
- Monosodium\_glutamate
- Pg\_tips
- Hamilton\_beach
- Trader\_joe

### Higher Question Use:

- Trans\_fats

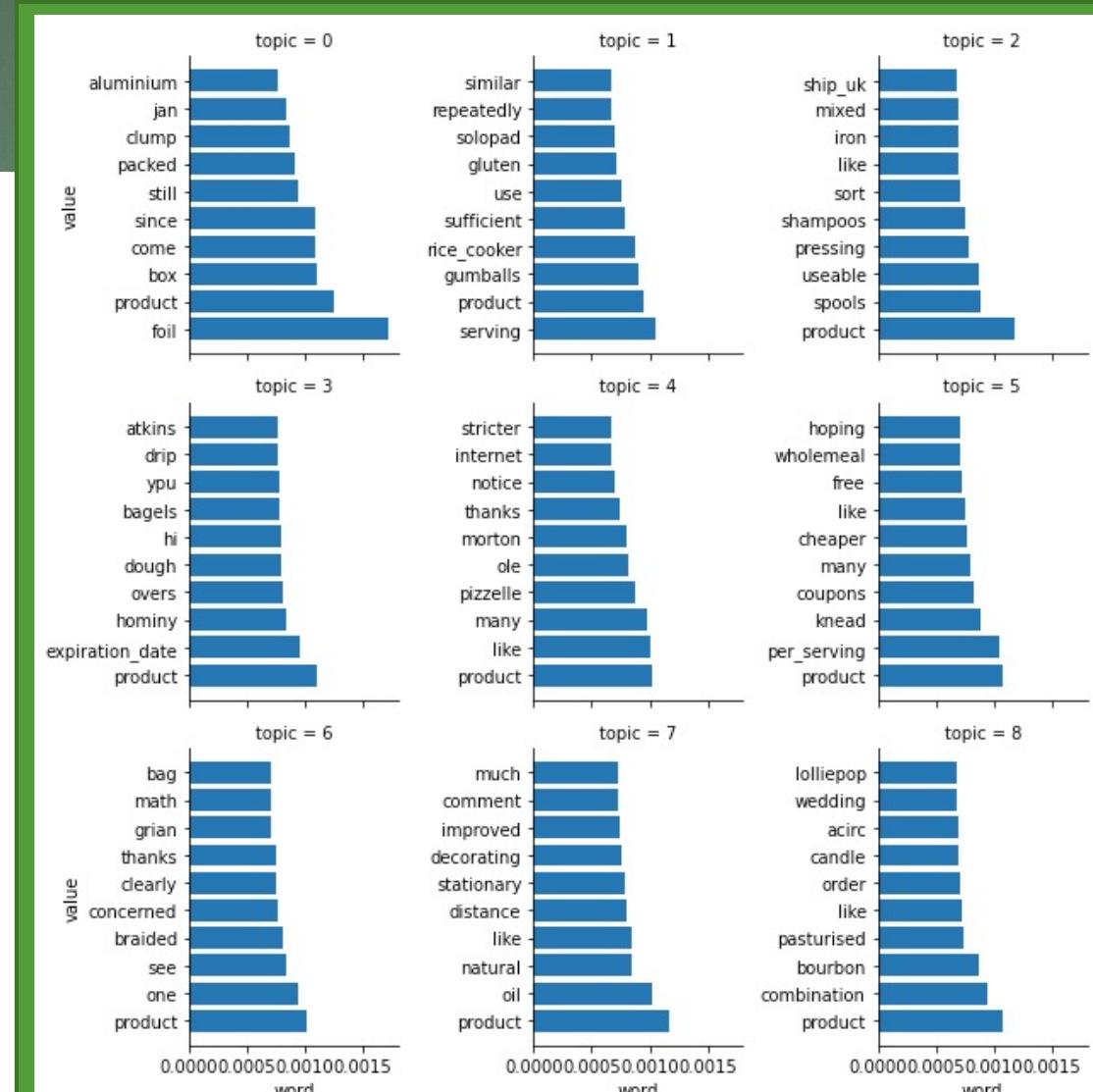
# Topic Modeling (Gensim)

Unstructured Data Predicted Categories

## Question Topics

- 0) Product understanding
- 1) Product match to equipment
- 2) product quality or clarity
- 3) Product shipment and conditions
- 4) Customer complaints on product use
- 5) Quantity and Sizing Understanding
- 6) Product sourcing conditions
- 7) Customer quantity concerns
- 8) Customer complaints on product use

*Analysis looked at larger grouping but found significance in these 9 topics for question words*

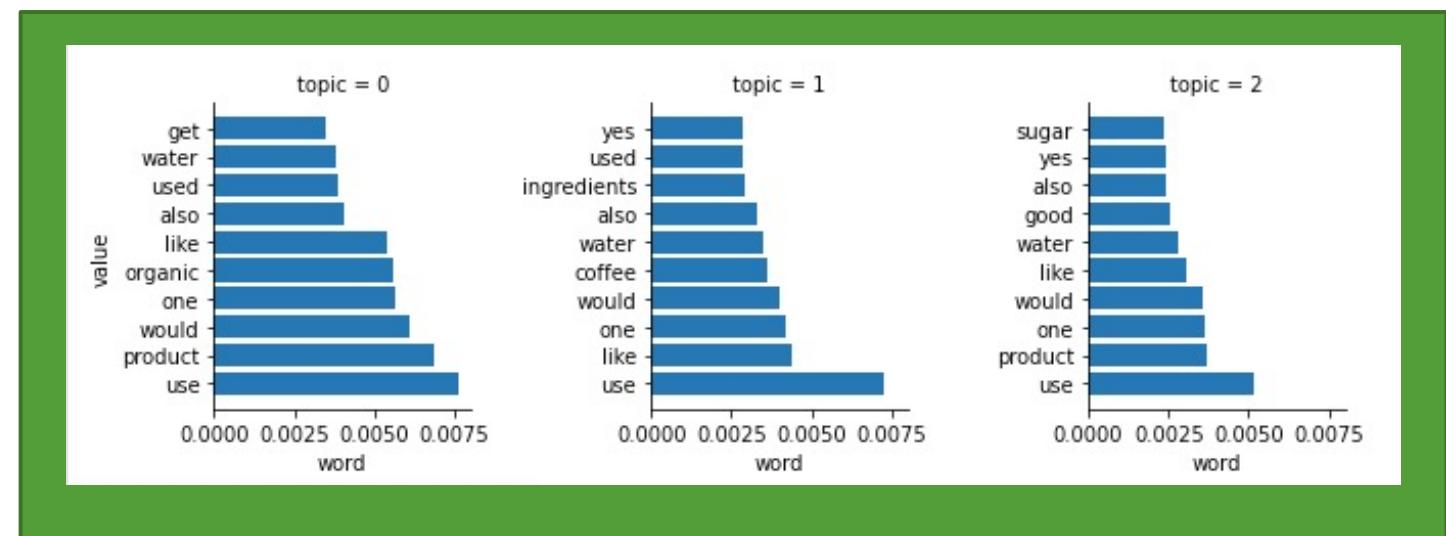


# Topic Modeling (Gensim)

Unstructured Data Predicted Categories

## Answer Topics

- 0) Product use instructions
- 1) Product Ingredients Detailed listing
- 2) Personal descriptions of product or competitor use



*Analysis looked at larger grouping but found significance in these 3 topics for answer words*

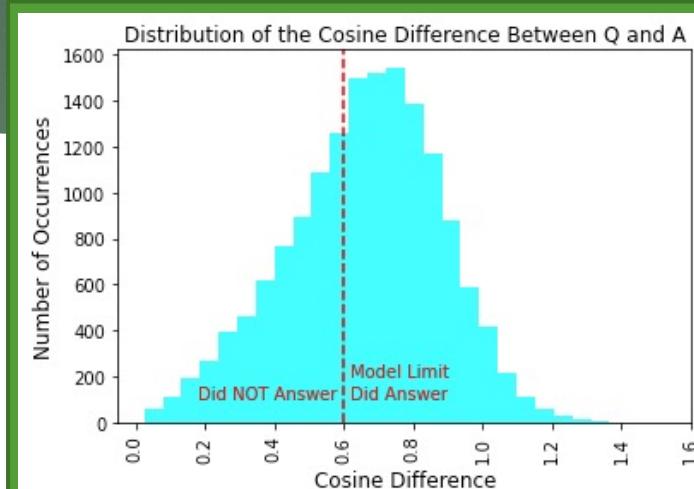
# Cosine Difference

Similarity of Q and A vectors measured by Cosine of Angle between them (Highest is Most Similar)

## Modeling

- 1) Gensim Doc2Vec
- 2) Vectorized Questions and Answers
- 3) Cosine difference between Q & A
- 4) Set model limit to 0.60

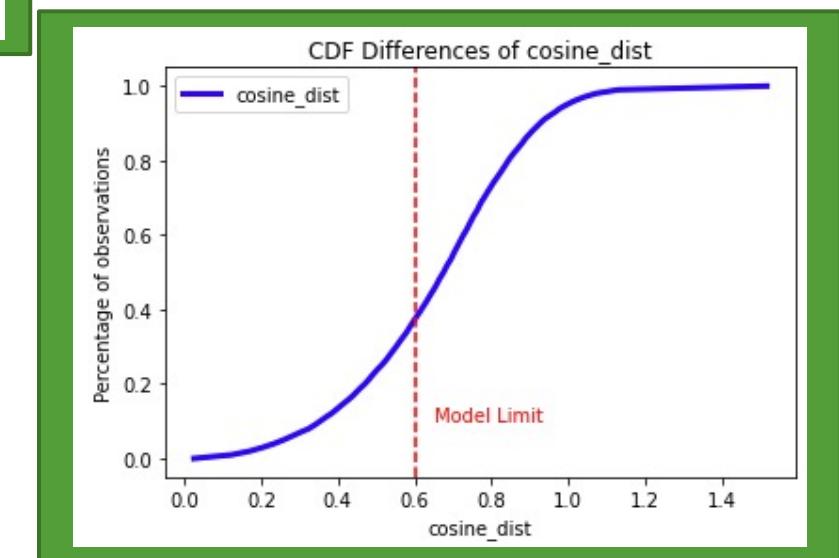
```
Cosine Distances described:  
count    15571.000000  
mean      0.657192  
std       0.221049  
min       0.024338  
25%      0.512988  
50%      0.673933  
75%      0.811545  
max      1.526258
```



*Applying this model predicts that only 40% of the questions were answered*

Interactive review of questions and associated answers best guess for model limit of 0.6 or higher for predicting actual answers to questions

The CDF shows the cumulative results that match the answer criteria based on a Cosine distance of 0.6 or below



# Levenshtein Distance

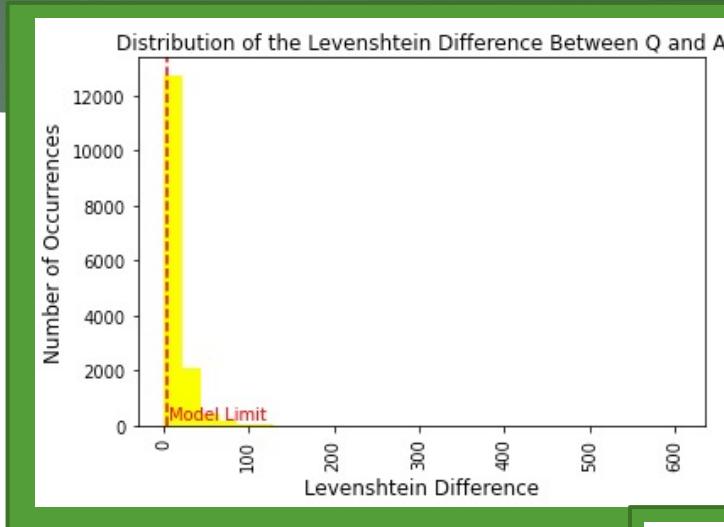
Similarity of Q & A vectors measured by cumulative change steps from Q words to A words (Lowest is Most Similar)

## Modeling

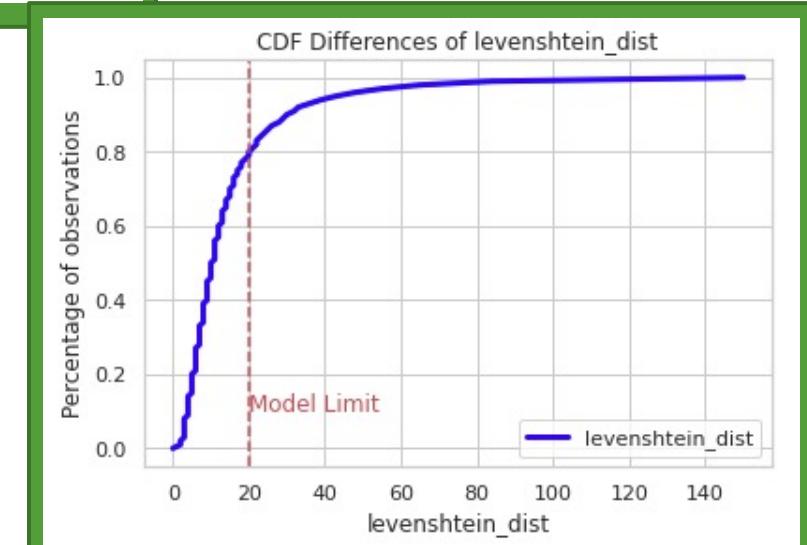
- 1) Levenshtein change difference
- 2) Determined word changes between Q & A
- 3) Noticed some Answers were restatements of Q but had only a few words difference. This was to add YES or NO to the question. These had low Cosine differences
- 4) Set model limit to 20

Levenshtein Distant Values:	
count	15571.000000
mean	15.150600
std	18.264013
min	0.000000
25%	6.000000
50%	10.000000
75%	17.000000
max	604.000000

Interactive review of questions and associated answers best guess for model limit of 20 or lower for predicting actual answers to questions



*Applying this model predicts that 80% of the questions were answered*



# BERT Sentence Similarity Modeling (torch)

Bidirectional Encoder Representations from Transformers is designed to pre-train from unlabeled text by jointly conditioning on both left and right context (Sentences as the Question and the Answer)

## BERT Model

### 1) Categorized Results:

Entailment: The answer aligned with the question and Agreed

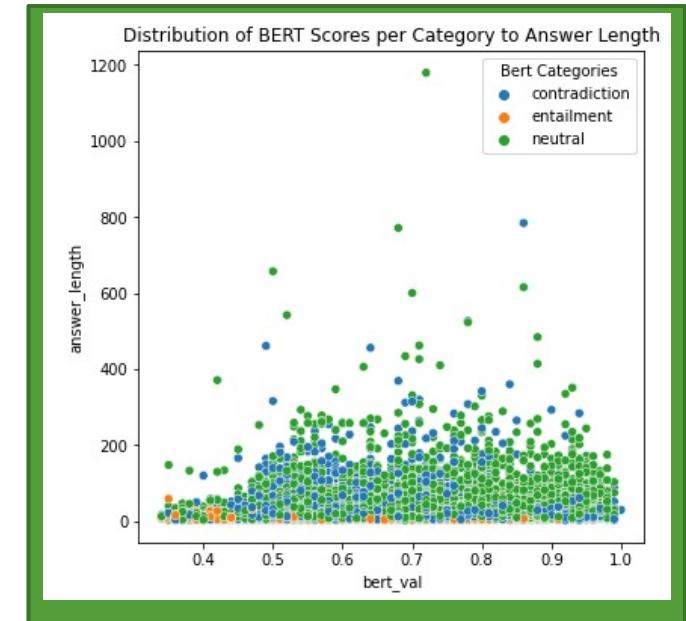
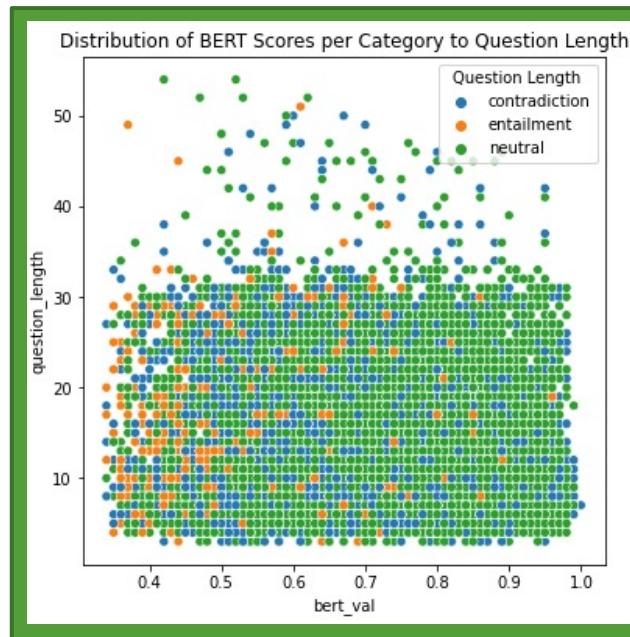
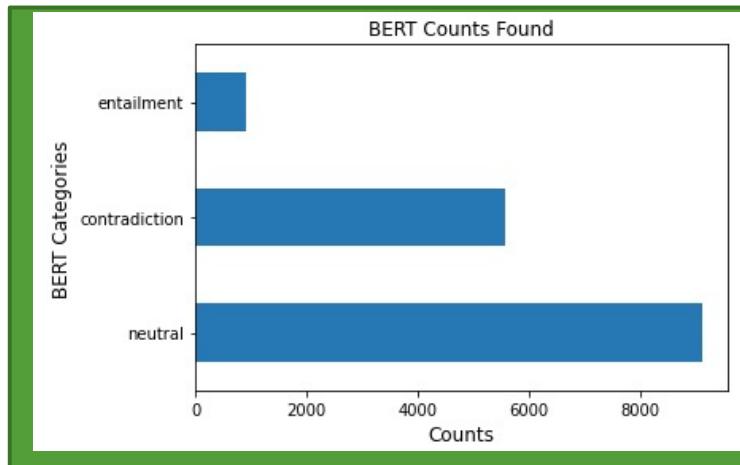
Contradiction: The answer aligned with the question and Disagreed

Neutral: The answer had no alignment with the question

### 2) 5% of answers Agreed

35% of answers Disagreed

58% Neutral



**Applying this model predicts that  
58.5% of the questions were  
categorized as Neutral – Neither  
answered positively or negatively**

# Label Answer Results (without Question context)

The data has no true labels to verify questions received answers.

Perform basic analysis and apply labels to answers viewed independently

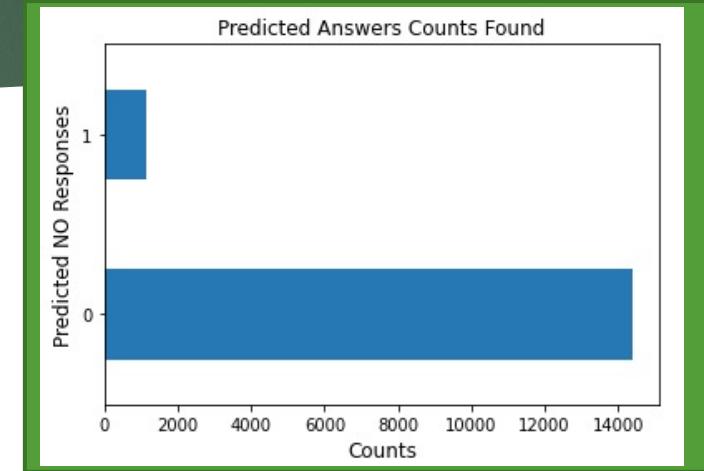
## Non Contextual Labeling of Answers as "Predicted Not Answered"

Look for these terms within the answer:

[ "don't know", "guessing", "do not know", "not sure", "no clue", "no idea", "never tried it", "never used it" ]

Set labels on these conditions alone and compare to model results

**Applying all previous models to the predicted NO answers shows poor alignment on True Negative detection. All less than 50% for negative detection.  
BERT was best at 45%**



Cosine Dist Model Confusion Matrix

		Actual
		-
Actual	-	770
	+	9003
Actual	+	391
	-	5407

Levenshtein Dist Model Confusion Matrix

		Actual
		-
Actual	-	853
	+	11423
Actual	+	283
	-	2749

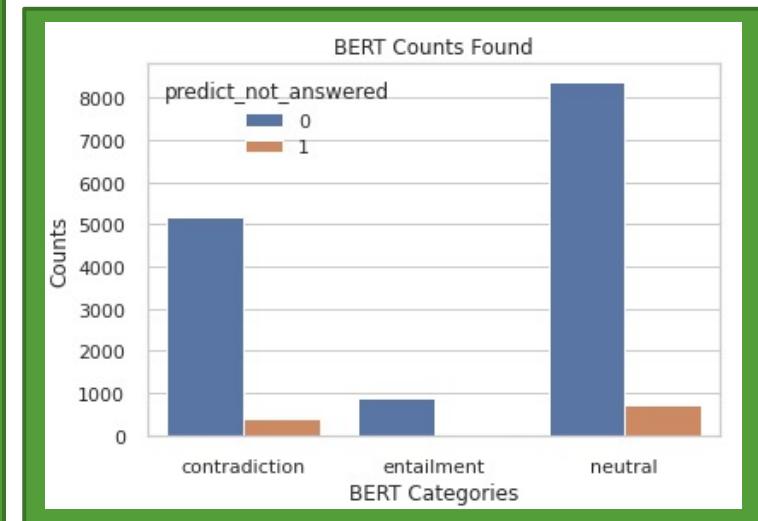
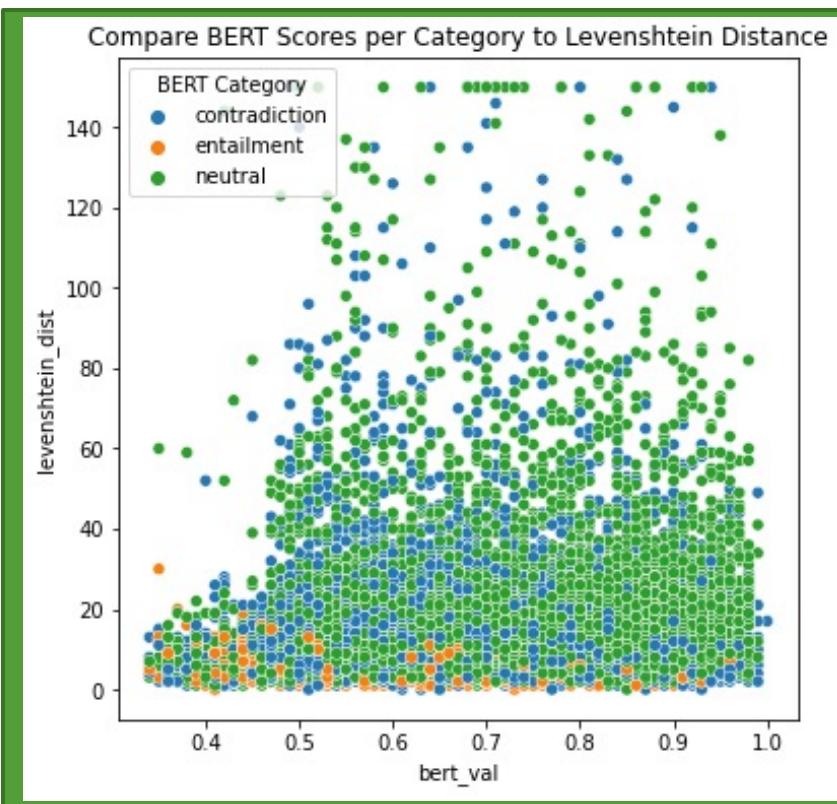
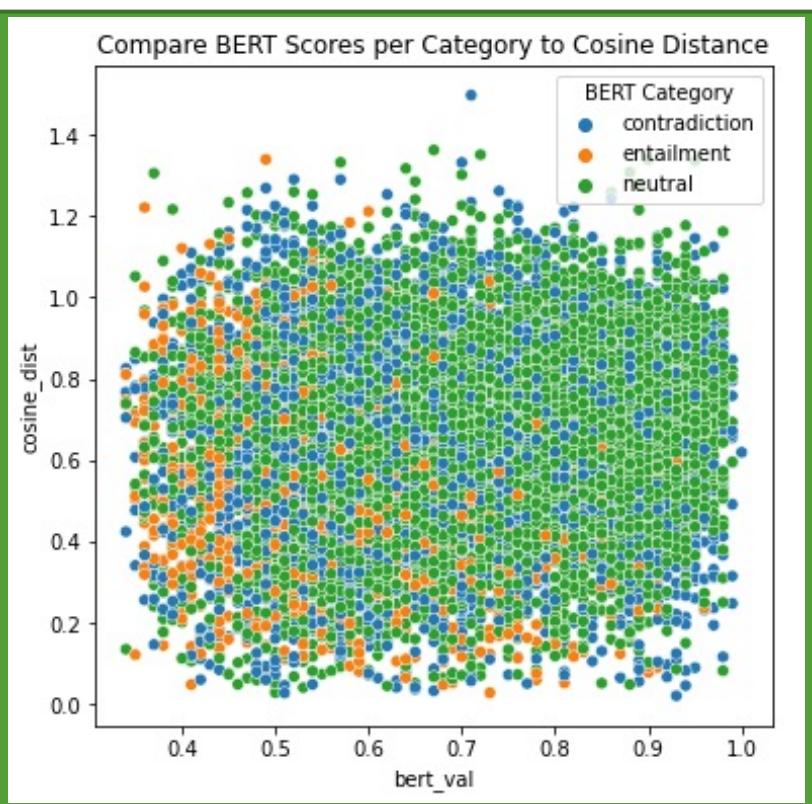
BERT Value Model Confusion Matrix

		Actual
		-
Actual	-	631
	+	7586
Actual	+	507
	-	6483

# Comparing the Models

Compare Cosine and Levenshtein to BERT categories

- Pattern with Entailment category: High identification of valid answer
- Contradiction and Neutral not as valid correlation

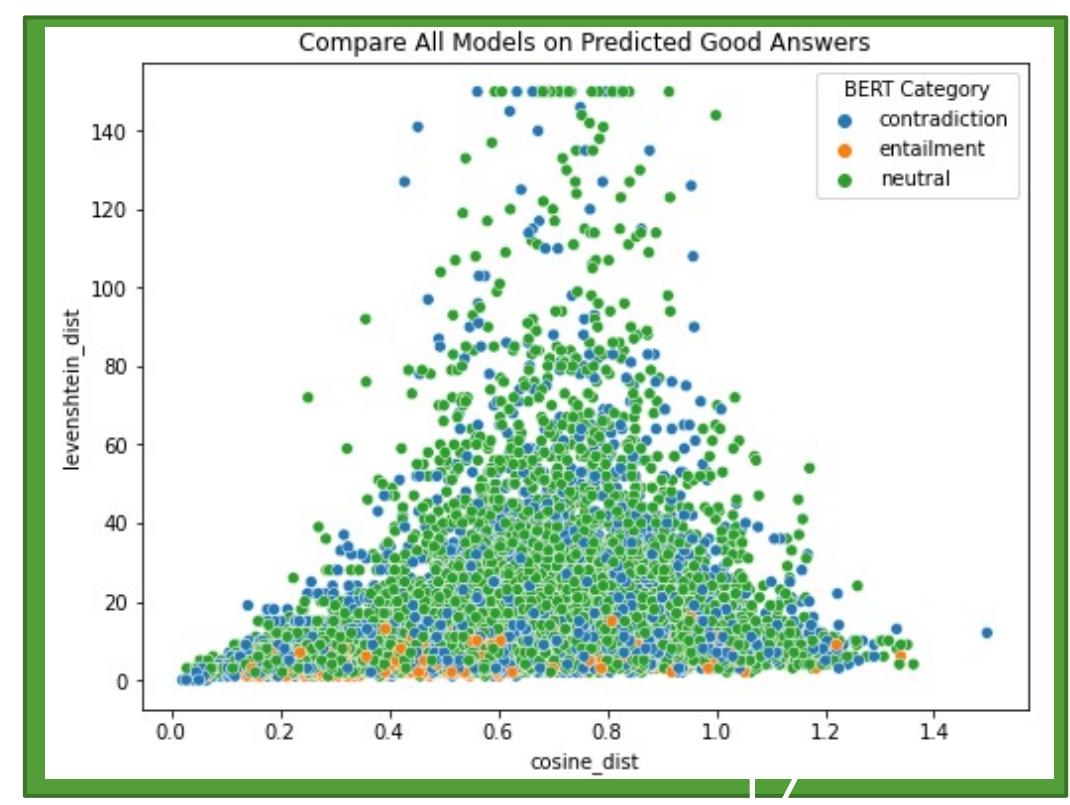
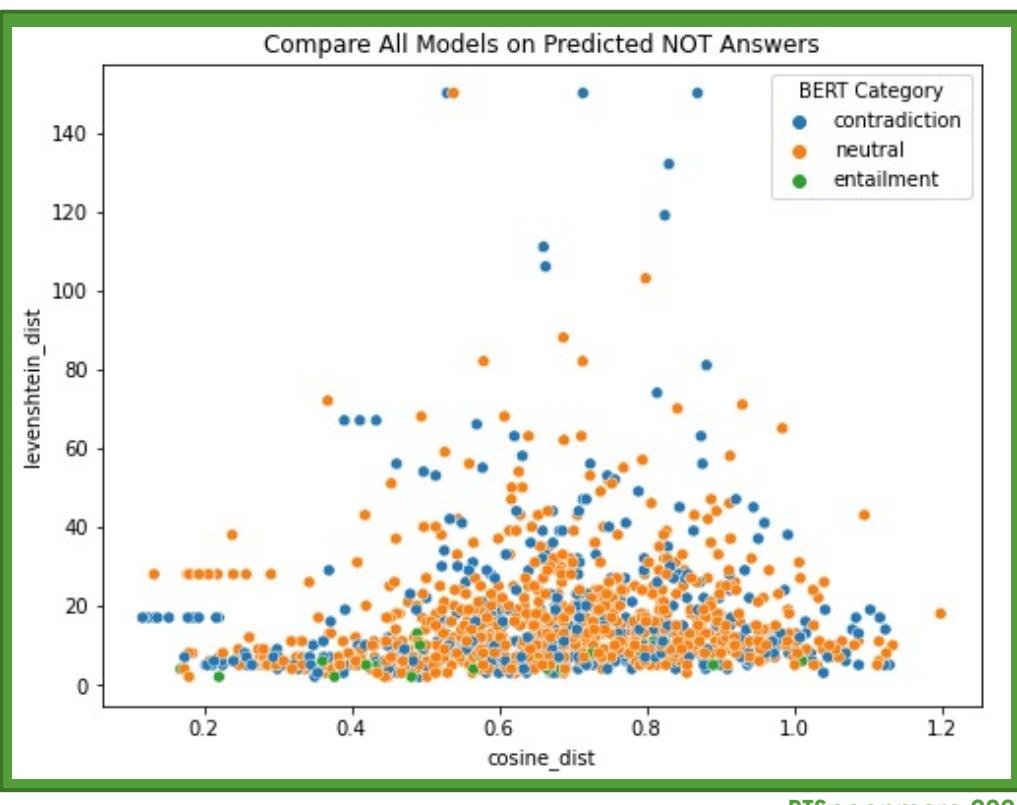


Predicted NOT answered fell between contraction and neutral BERT categories

# Comparing the Models

Compare Cosine and Levenshtein to BERT categories ONLY using the predicted NOT answers

- Neutral answers higher alignment with predicted NOT answers
- Some Contradiction categories in NOT, but few Entailment



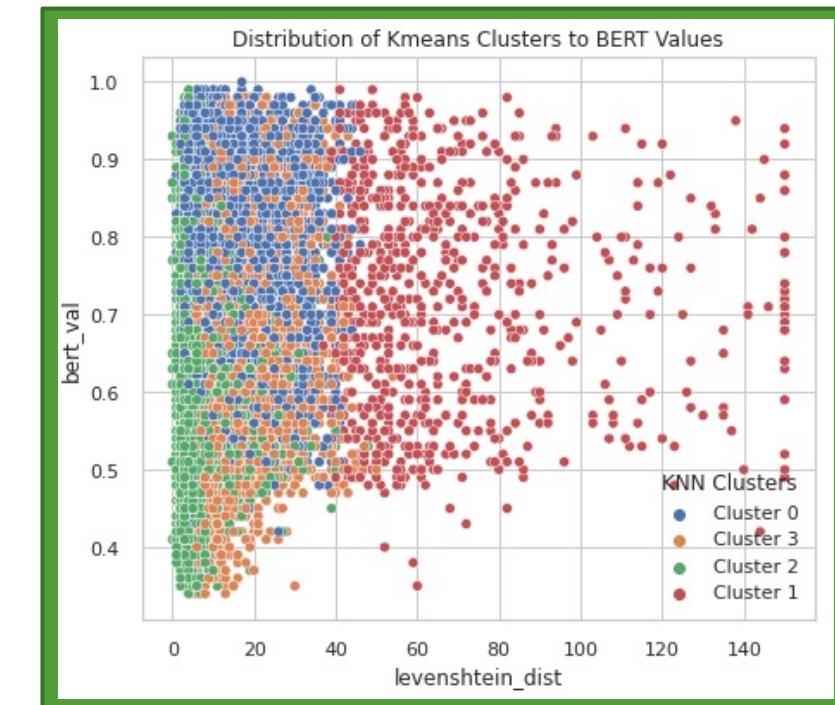
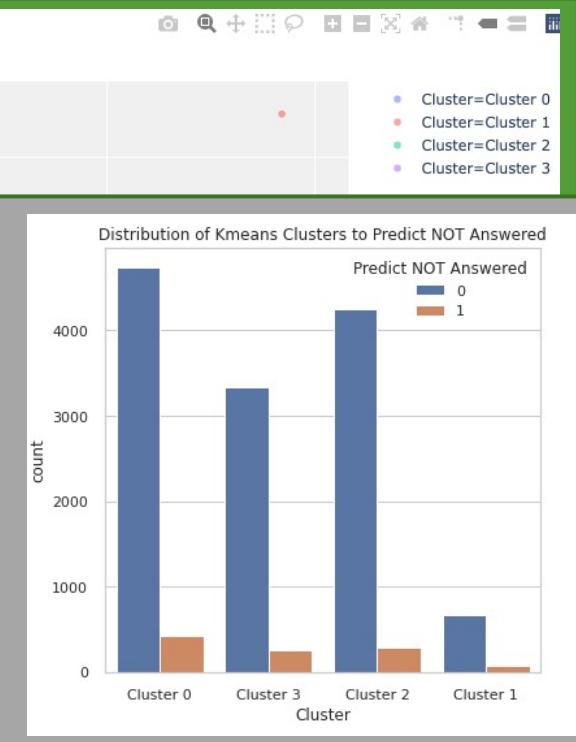
# Clustering and Regression on Model Outputs

Utilized PYCARET to search for patterns on model outputs

Multiple Regression models did not find a  $R^2$  above 0.011

Kmeans clustering optimized at 5 Clusters but no strong correlation found to predicted NOT answers

2D Cluster PCA Plot



# Learning from Model Comparison

- Answering questions requires context. Analysis of questions or answers individually does not provide enough context to provide dependable results in an unlabeled dataset.
- The BERT Model did highlight 5.8% of the answers as dependably valid, 35.8% as a contradiction (most Negative responses to questions, leaving 58.4% as Neutral answers. *BERT can be used to decrease the dataset down to 58.4% to reduce the volume of answers to review*
- Direct text searches of answers identified 8% predictable NOT answered questions. This did not enable Kmeans analysis to find similar NOT answers.

# Recommendations

Utilize Topic Modeling for Next Steps

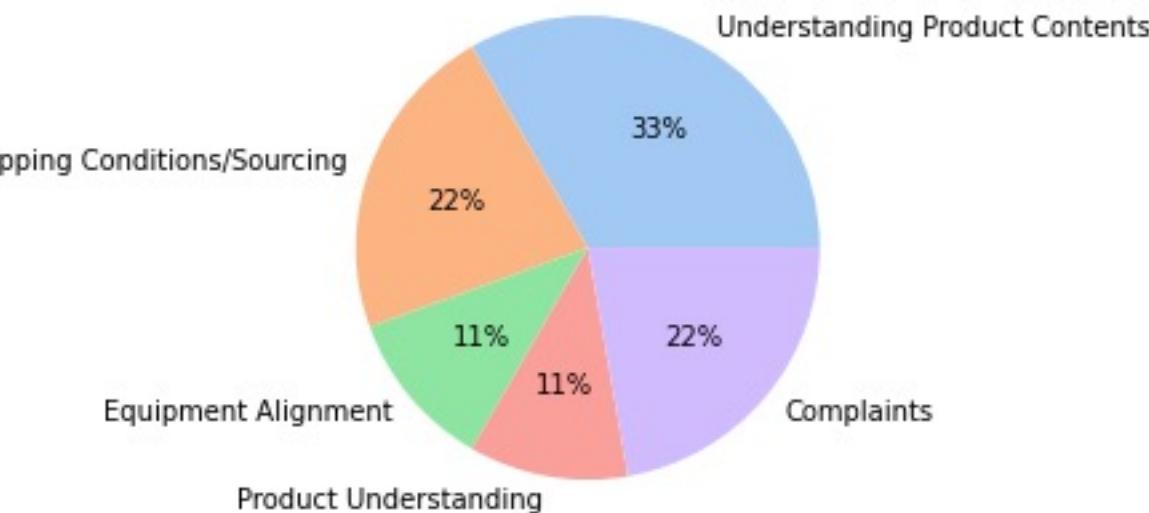
## 1) Preventable Questions (67%)

- Understanding Product Contents (List contents, allergens, organics, health details)
- Shipping Conditions or Sourcing (State Origin, Refrigeration)
- Equipment Alignment (State what products fit)

## 2) Questions Requiring Response (33%)

- Product Understanding (Respond directly to how to use, combine, apply product)
- Complaints (Respond directly to customer and specific concerns)

Breakdown of Answer Topic Areas for Amazon



# Future Model Recommendations

Since Amazon has access to the contextual details on product basis, apply a BERT model for Question context with Answer verification which has been purported to achieve a precision of 76.2% and a recall of 57.9% (Culberg)

Culberg, Kevin. Question Answering with Bert and Answer Verification. - Stanford University. Stanford University, <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15763476.pdf>.

# What I Learned from the Process

Reflections on Personal Lessons Learned

- 1) The analysis tools struggle with 1 & 2 word Questions and Answers: Removed from Model
- 2) Because unsupervised data, Each step required personal data review of Q&A alignments
- 3) Difference between Word2Vec and Doc2Vec models. Utilized Doc2Vec which required question bag of words to be in form of lists of lists
- 4) Bag of Words analysis required Questions and Answers to be in a flat file format
- 5) Maintained processed Question and Answer lists in list format to expedite processing time, as initial approach to build into dataframe memory structure overloaded my computer
- 6) Corpus is a list of indicies and counts, Dictionary is a list of indicies and word locations. Had to keep clear for each model to align with results to see how it related to each word
- 7) Visualizations of counts and scatter plots explain relationships much better than tables
- 8) Topic Modeling very manual: required review of all questions topic words to determine topic
- 9) There is a difference between Cosine Angle, Soft Cosine, and Cosine difference.
- 10) Levenshtein is very sensitive to text length, so long answers had big impact on measure
- 11) Cosine alone did not model dynamics completely
- 12) Bert Model had heavy system requirements and crashed my computer multiple times

# Output Discoveries from the Process

Applied Learning from the Results

- 1) The Amazon data challenging: Prank questions, non serious answers, jargon, inconsistent slang, inconsistent styles, answers conversation style
- 2) A pretrained model failed to predict the results. First pass used Google trained model, but over 12,000 unique words in Amazon data were not found. Too many abbreviations and jargon
- 3) Some Answers just restated the Question with adding YES or NO, thus the levenshtein model was needed.
- 4) The question topics categories predict areas for improvement to lower number of question types
- 5) BERT model required heavy computing resources (20 minutes for main step to execute)
- 6) Context is needed to truly evaluate meaning of Q & A