NLP Analysis of Amazon Question and Answer banks

HTTPS://WWW.BIZJOURNALS.COM/COLUMBUS/NEWS/2017/08/03/REVIEWING-AMAZON-RESTAURANTS.HTML

CAPSTONE PROJECT

BOB SPOONMORE

OCT 2021

Problem Identification:

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

HTTPS://WWW.AMAZON.COM/FMC/LEARN-MORE?REF_=PRIMENOW

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

A Capstone project:

- Applying NLP modeling to Amazon Question and Answer banks
- Analyzing Unsupervised and Unstructured Text Fields
- Data: 2018 Data Extraction for Amazon category: Grocery and Fine Food
- Predicting the probability of a question being answered correctly
- Identifying the Categories of questions based on Topic Modeling
- Applying models of text analysis on answers, Cosine and Levenshtein Distances between questions and answers, and BERT sentence semantic modeling approaches

Cost of Not Answering Customer Questions

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:

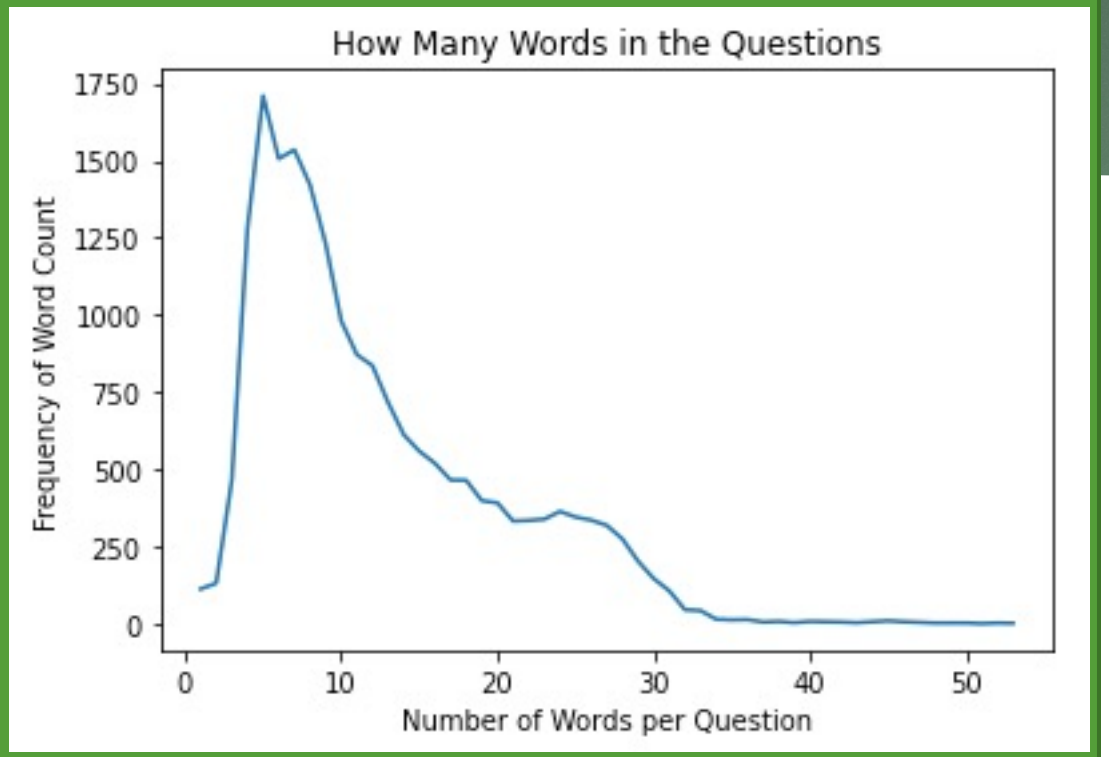
Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems
Mengting Wan, Julian McAuley International Conference on Data Mining (ICDM), 2016

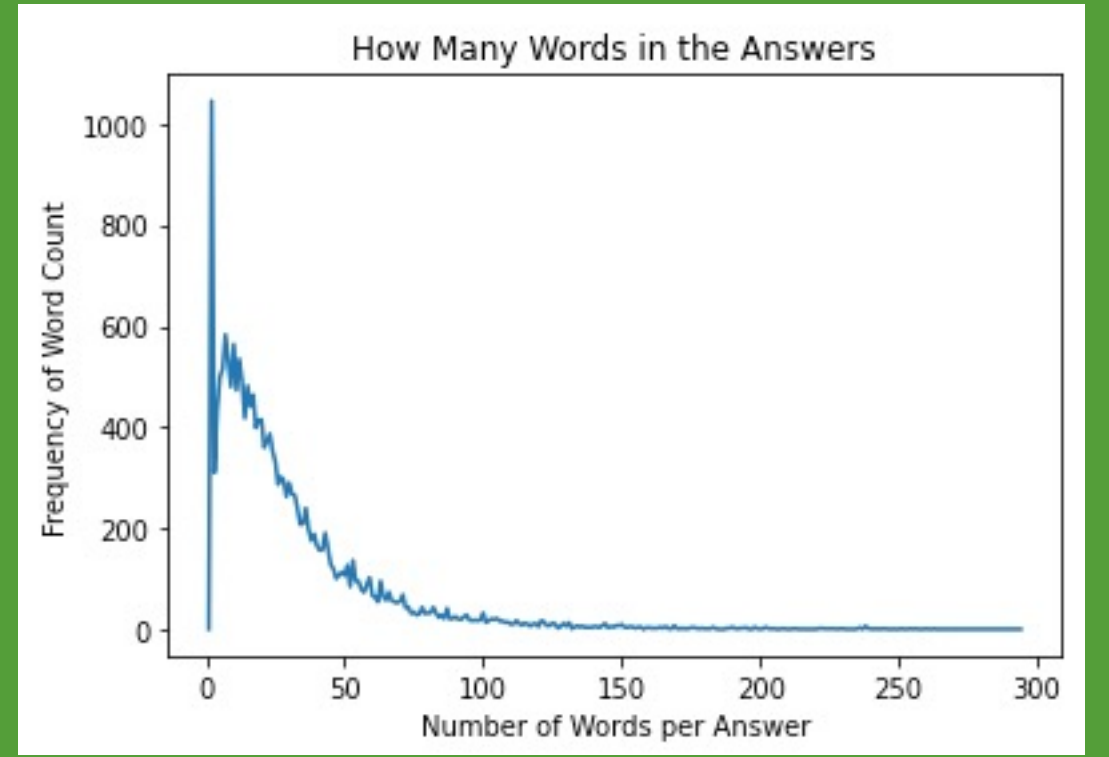
Addressing complex and subjective product-related queries with customer reviews Julian McAuley, Alex Yang World Wide Web (WWW), 2016

	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 yello...	I think that the yellow is the most mild. The ...	NaN
1	yes/no	9742356831	Apr 2, 2014	1.396422e+09	Is there MSG in it?	No MSG in Mae Ploy curry pastes.	N
2	open-ended	9742356831	Apr 5, 2015	1.428217e+09	what are the ingredients exactly in this produ...	The ingredients are listed in the description!	NaN
3	open-ended	9742356831	Aug 19, 2014	1.408432e+09	How important is the expiración date on...	I never pay attention to it myself. The ingred...	NaN
4	open-ended	9742356831	Aug 2, 2014	1.406963e+09	The product description says 14 oz., but the p...	We bought the 14oz for just under $5.	NaN

shape: (19538, 7)

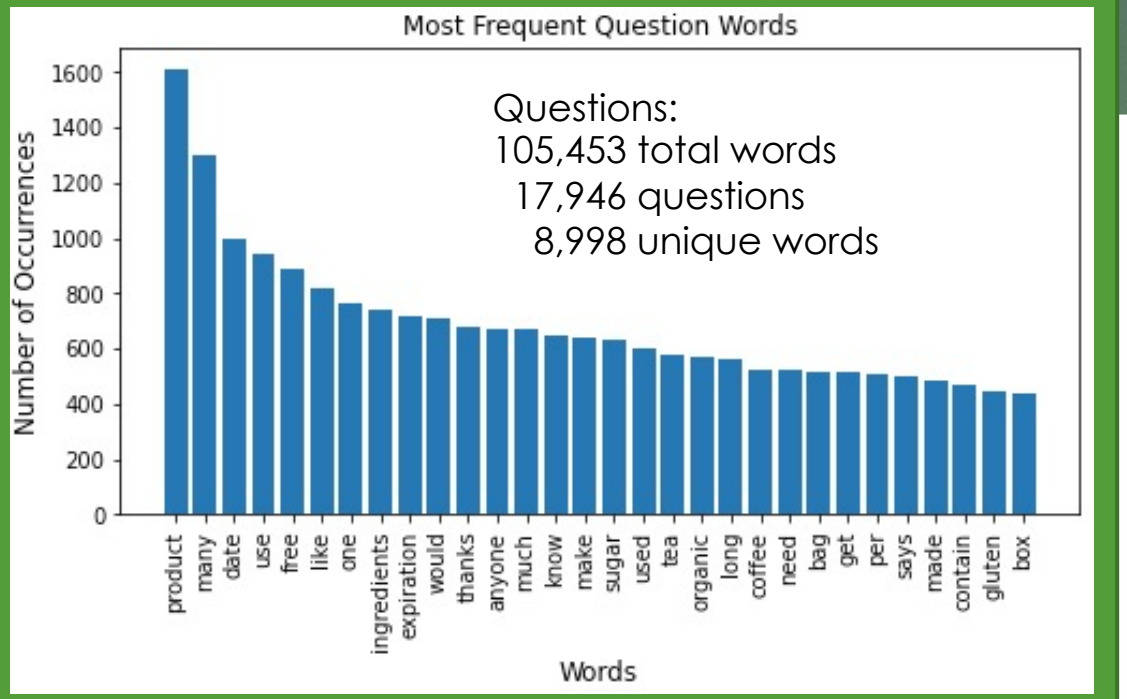
Exploratory Data Analysis

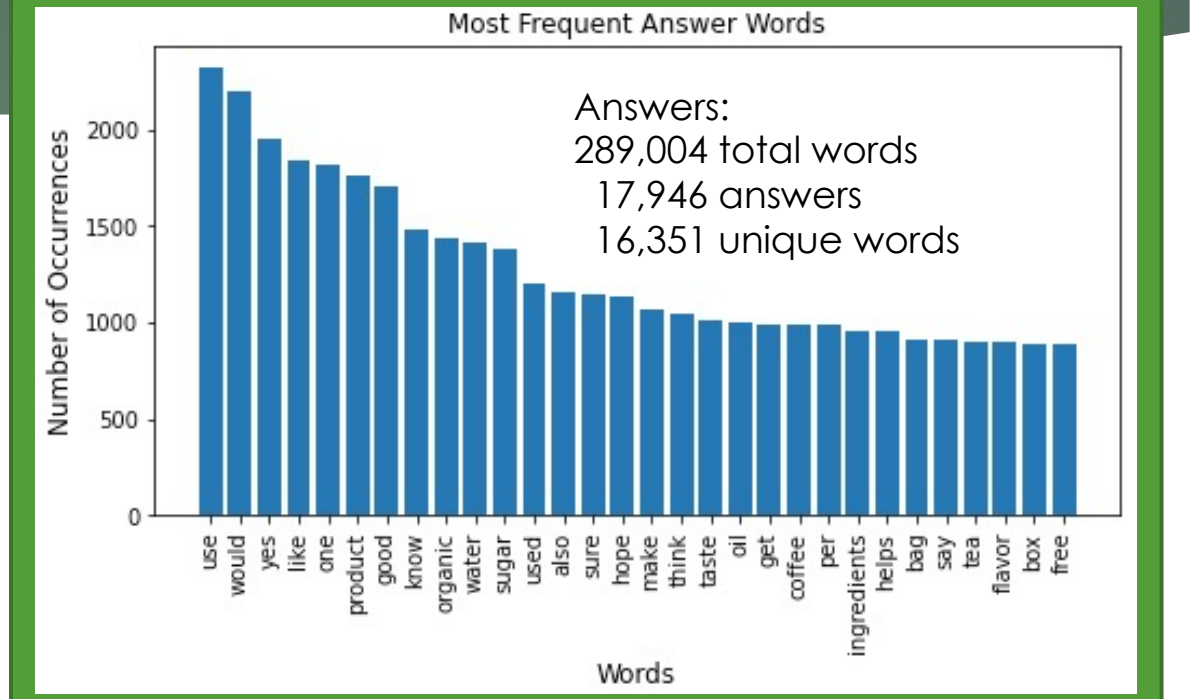




No null values found
Removed all 1 or 2 word questions and answers as these would bias models

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
- Maintained Integrity of questions as separate lists of words
- 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) Show top values each, similar to Bag of Words frequency but weighting adjusted

Questions

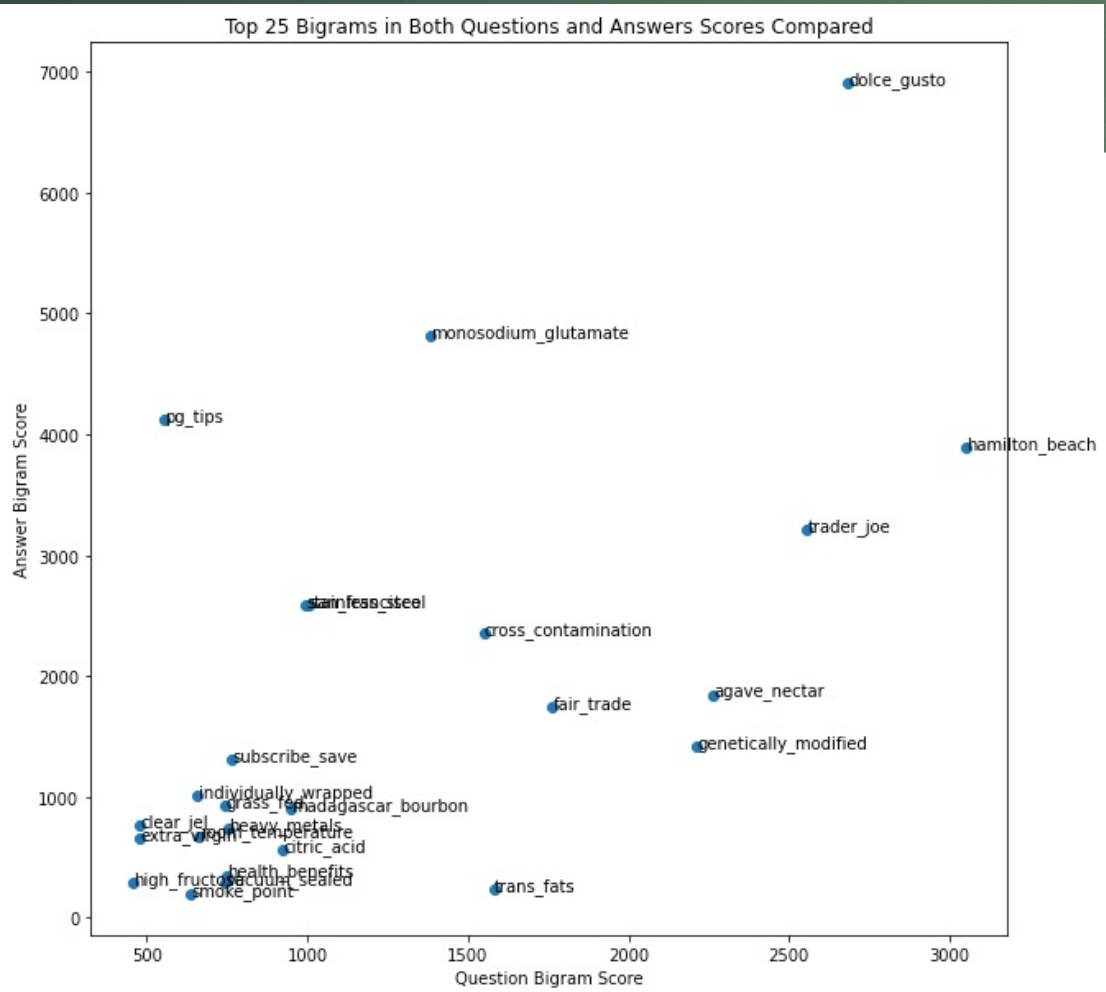
```
Top feature Names:
curry 0.5280719034995475
red 0.37460475940536125
yellow 0.4292527979030419
compared 0.4501747390819909
heat 0.4403363245862047
```

Answers

```
Top feature Names:
red 0.24963019415811116
profile 0.401166162876837 43
flavor 0.2004650839356598
deeper 0.427185427761358 87
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:

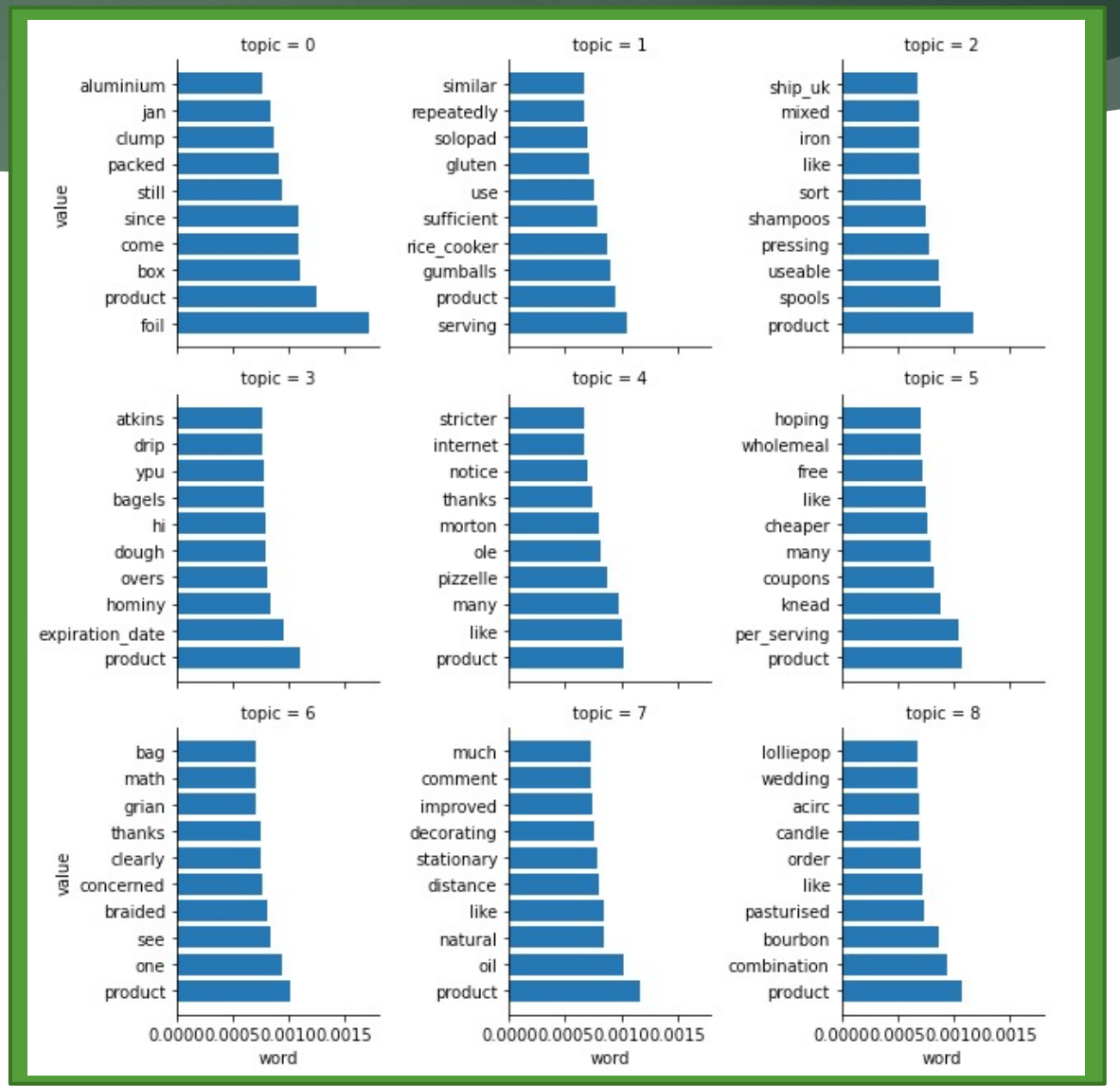
- Dolce_gusto
- Monosodium_glutamate
- Pg_tips
- Hamilton_beach
- Trader_joe

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

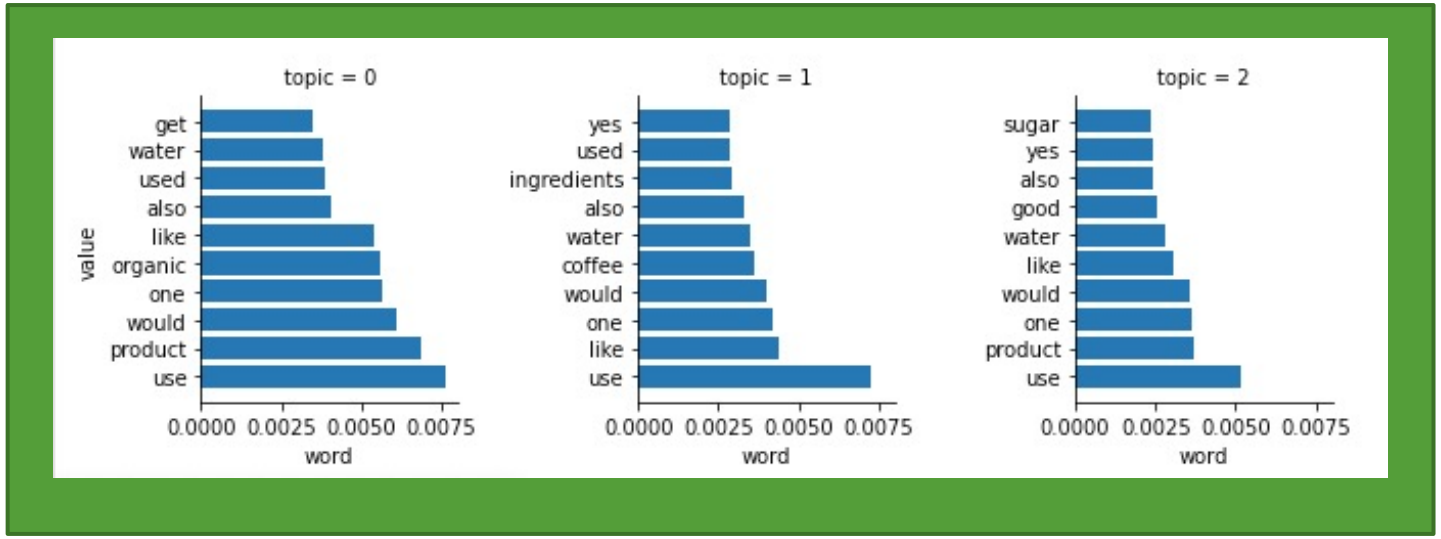
Analysis looked at larger grouping but found significance in these 9 topics for question words

Topic Modeling (Gensim)

Unstructured Data Predicted Categories

Answer Topics

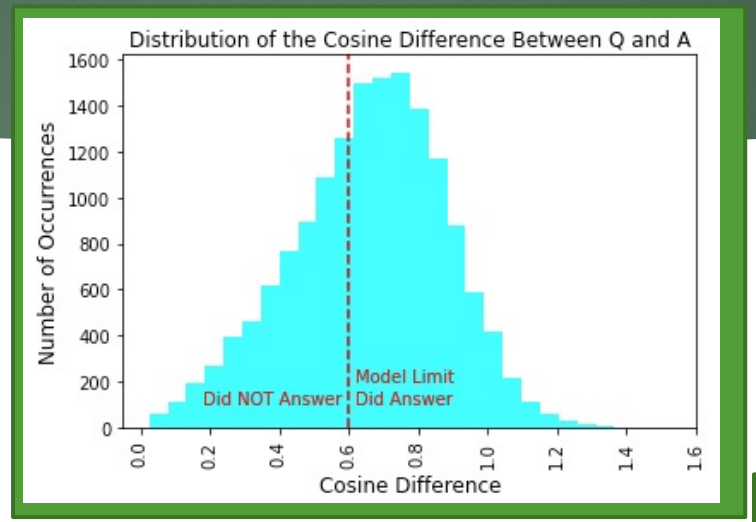
0) Product use instructions
1) Product Ingredients Detailed listing
2) Personal descriptions of product or competitor use



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

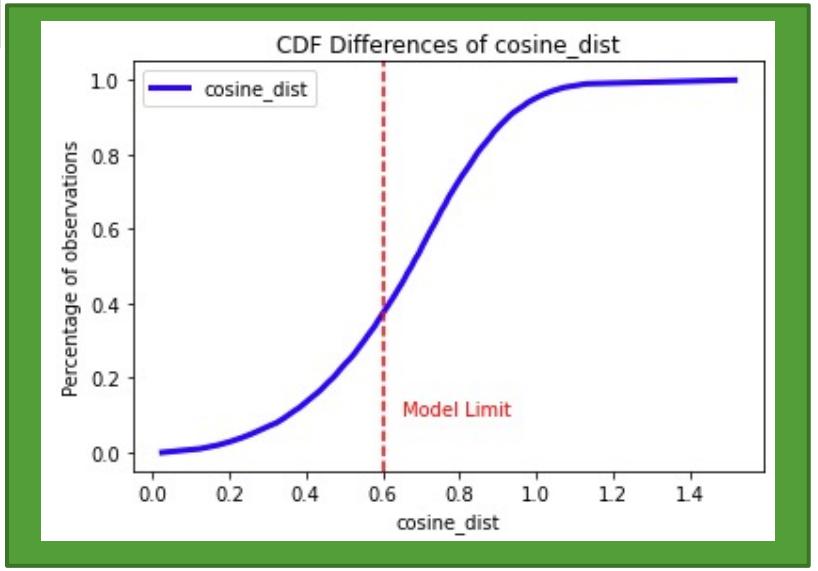
1) Gensim Doc2Vec
2) Vectorized Questions and Answers
3) Cosine difference between Q & A
4) Set model limit to 0.60

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

Interactive review of questions and associated answers determined model limit of 0.6 or higher best predicted actual answers to questions

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

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

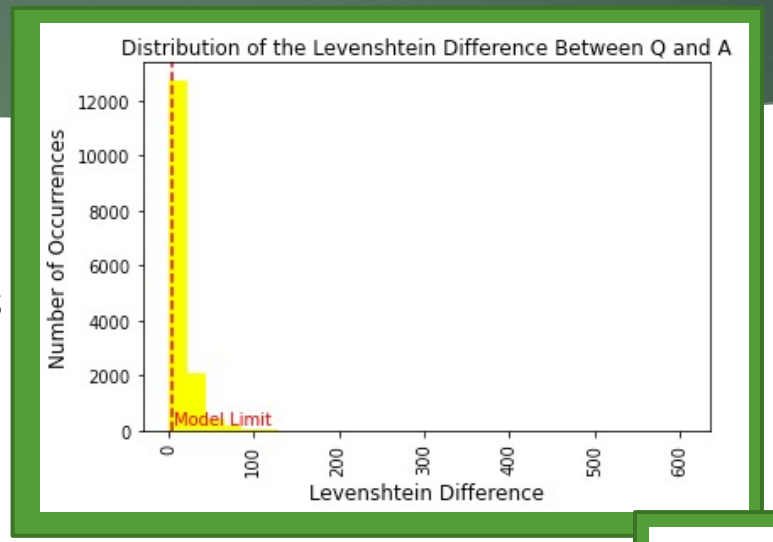


Levenshtein Distance

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

Modeling

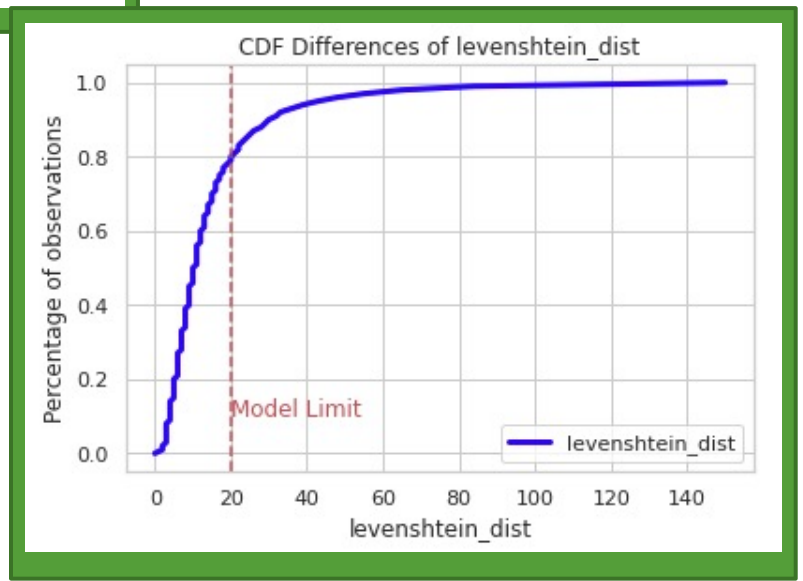
1) Levenshtein change difference
2) Determined word changes between Q & A
3) Noticed some Answers were restatements of Q but had only a few words difference. This was to add YES or NO to the question. These had low Cosine differences
4) Set model limit to 20



The CDF shows the cumulative results that match the answer criteria based on a Levenshtein distance of 20 or below

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

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



Interactive review of questions and associated answers determined model limit of 20 or lower best predicted actual answers to questions

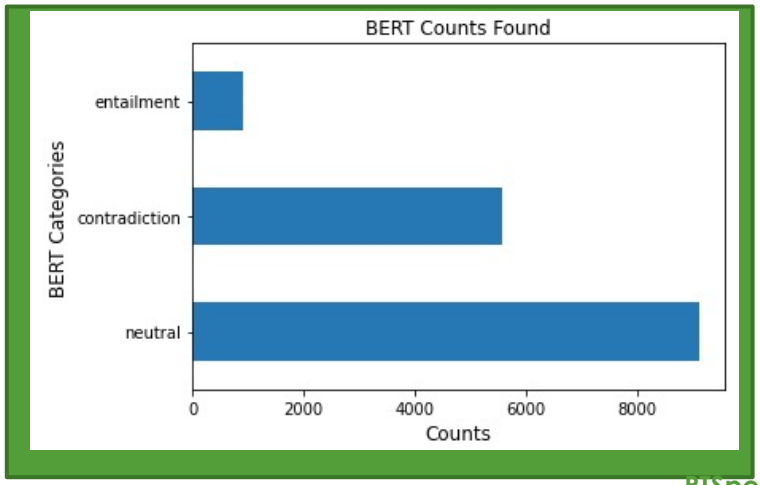
BERT Sentence Similarity Modeling (torch)

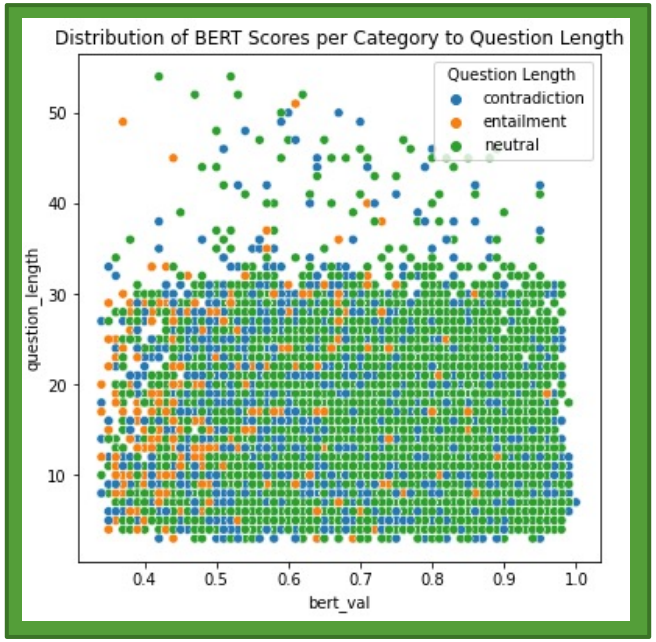
Bidirectional Encoder Representations from Transformers is designed to pre-train from unlabeled text by jointly conditioning on both left and right context (Sentences as the Question and the Answer)

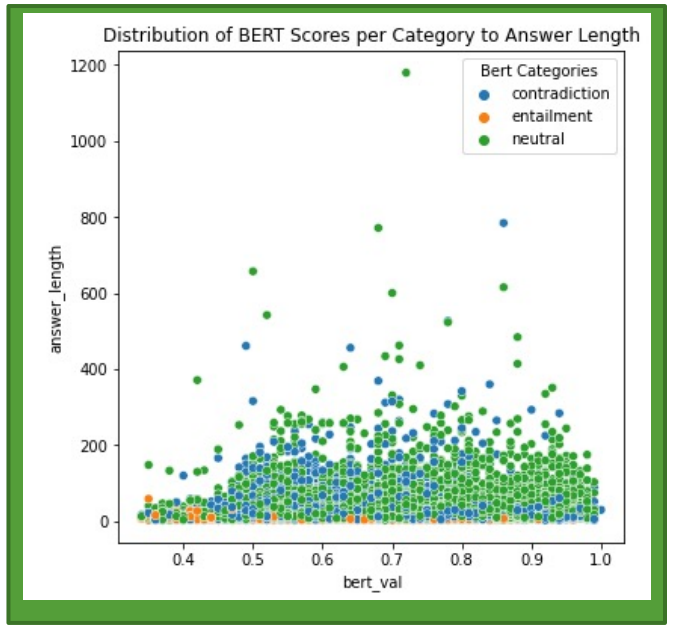
BERT Model

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



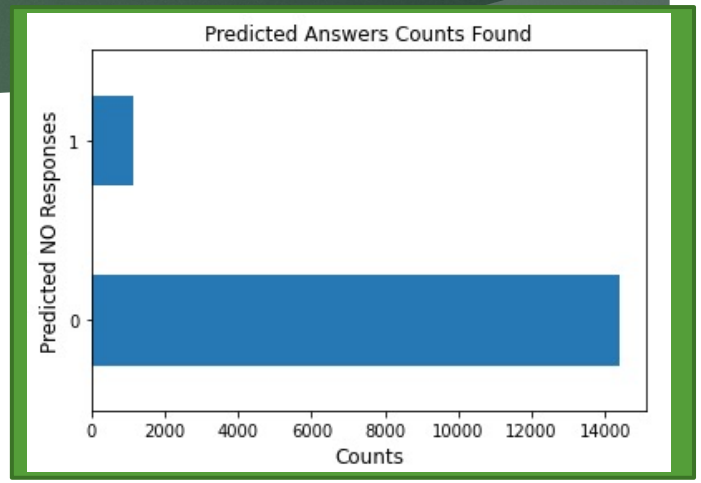




Applying this model predicts that 58.5% of the questions were categorized as Neutral – Neither answered positively or negatively

Label Answer Results (without Question context)

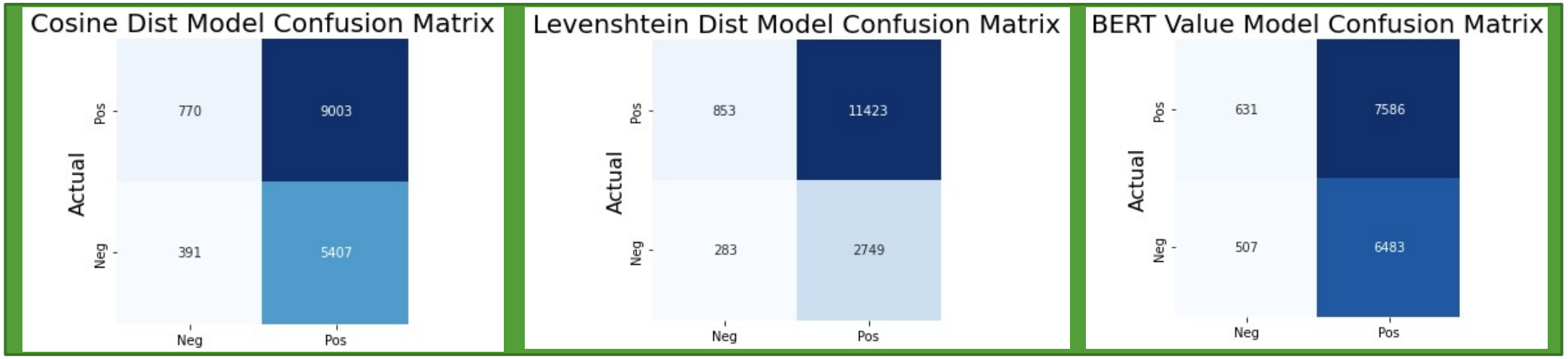
The data has no true labels to verify questions received answers.
Perform basic analysis and apply labels to answers viewed independently



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

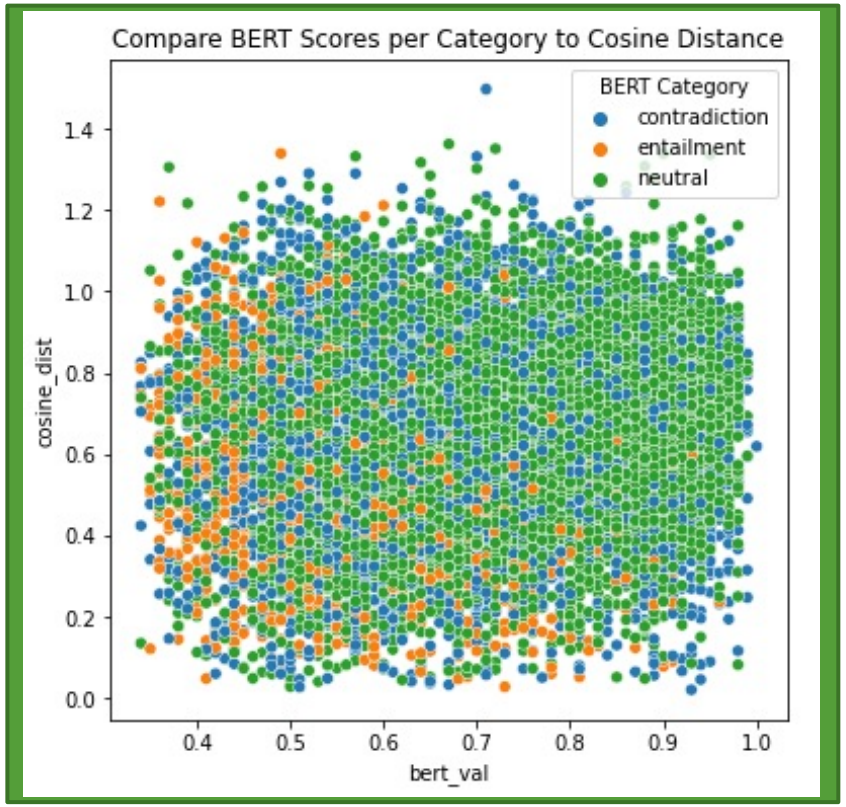
Applying all previous models to the predicted NO answers shows poor alignment on True Negative detection. All less than 50% for negative detection. BERT was best at 45%

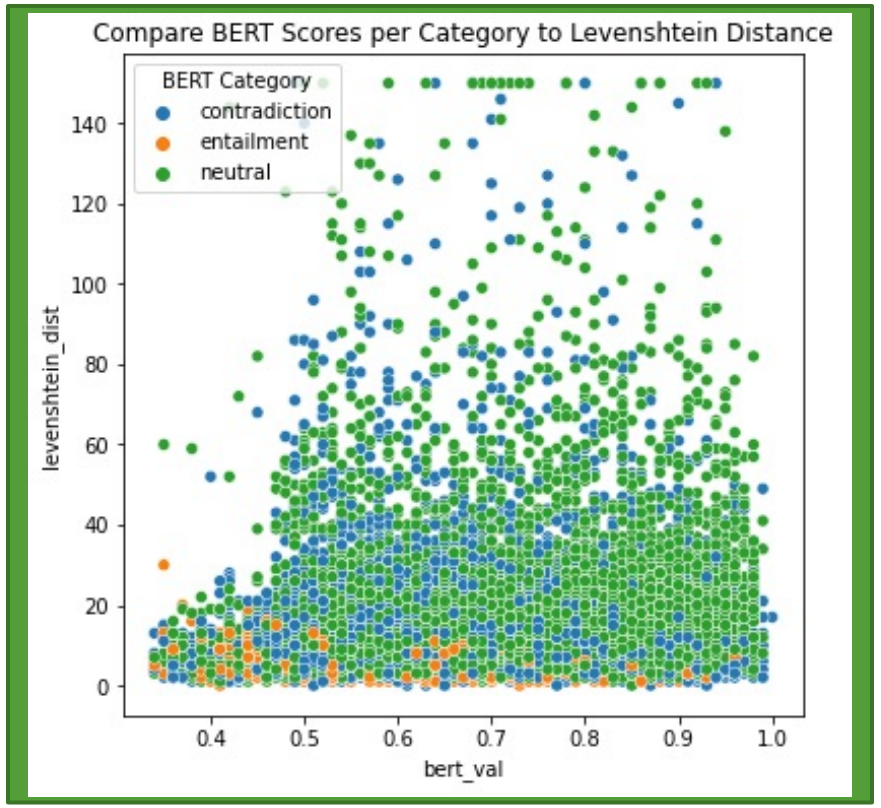


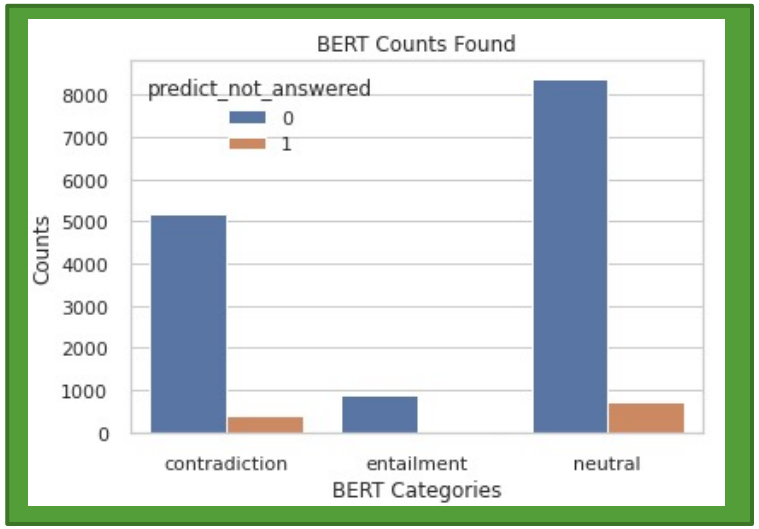
Comparing the Models

Compare Cosine and Levenshtein to BERT categories

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





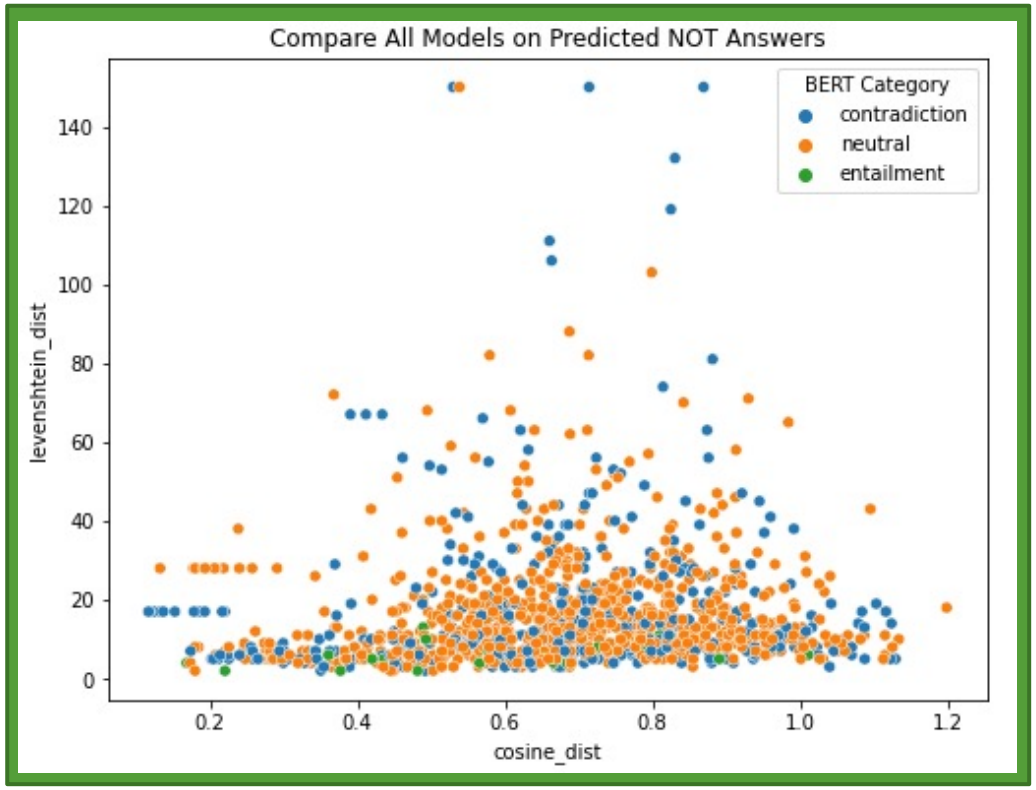


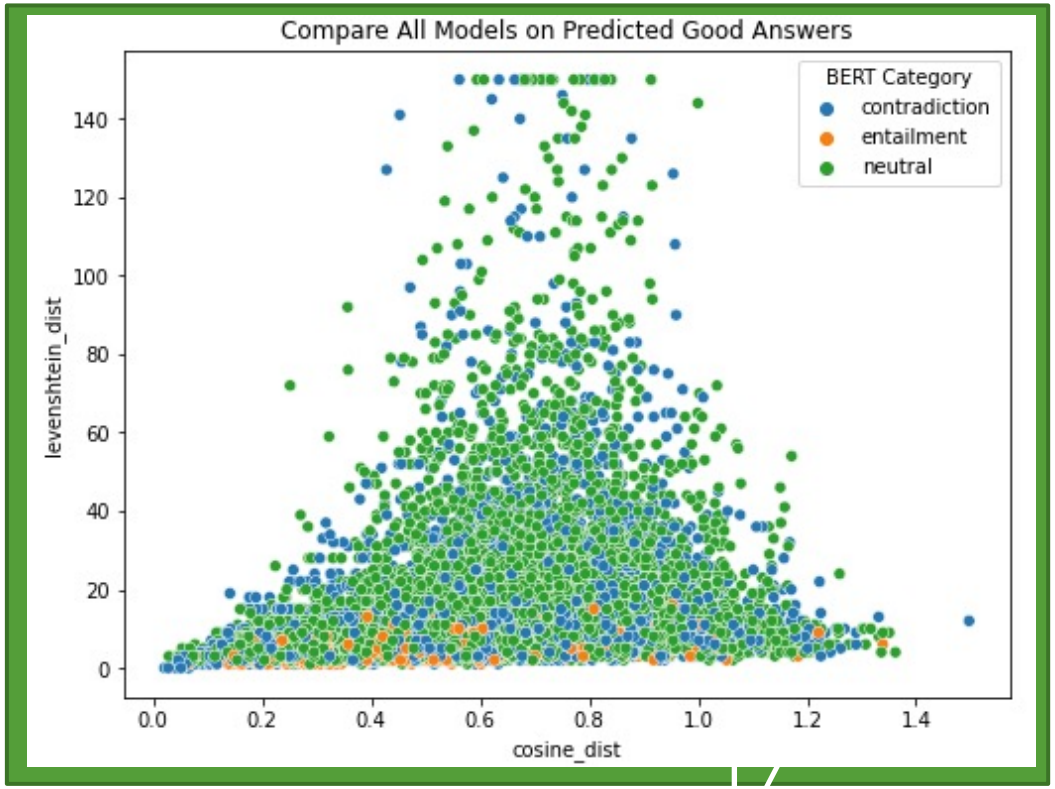
Predicted NOT answered fell between contraction and neutral BERT categories

Comparing the Models

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

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

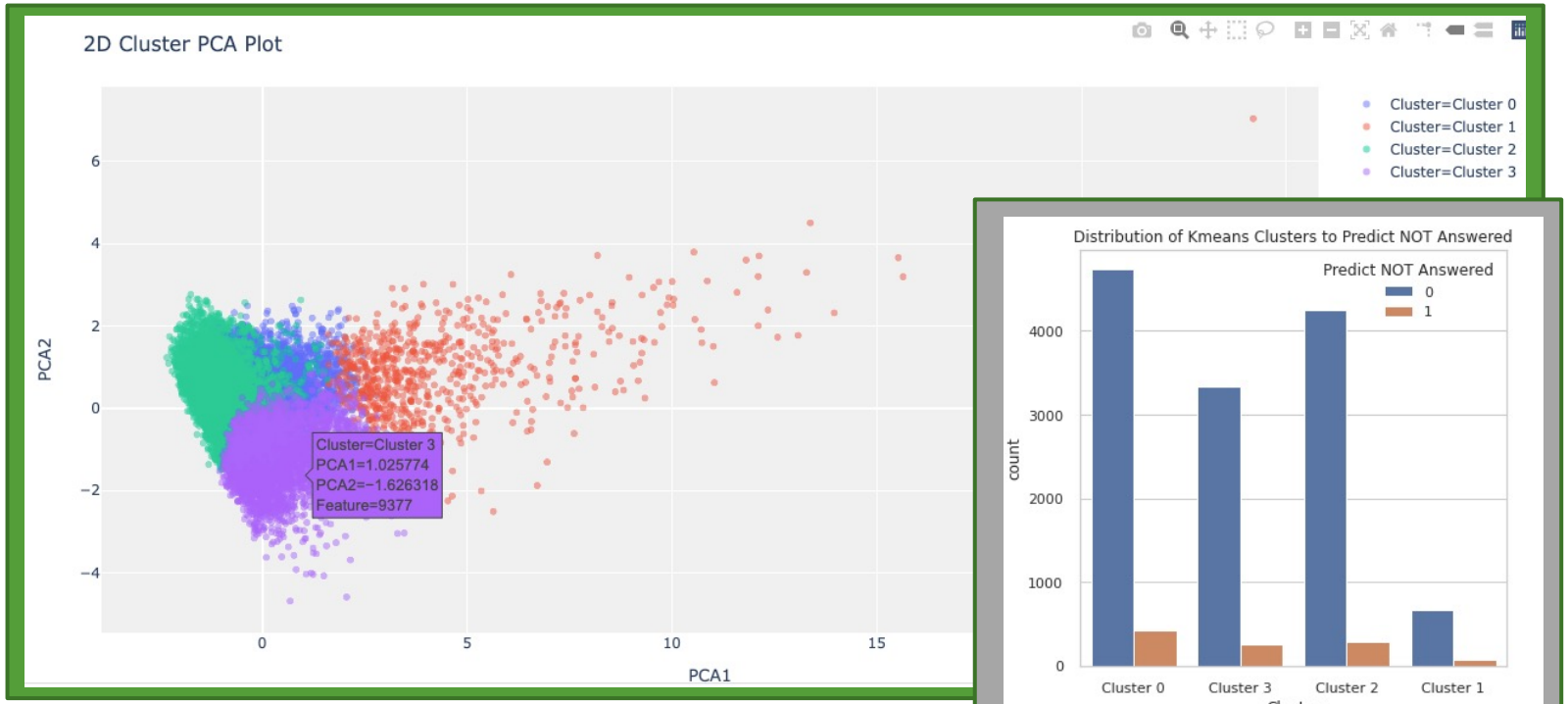


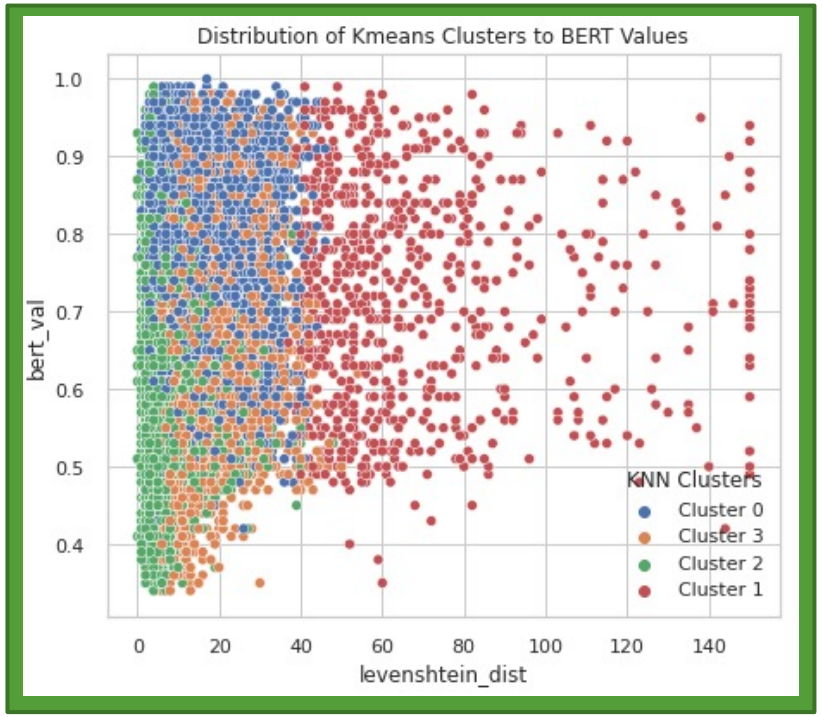


Clustering and Regression on Model Outputs

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





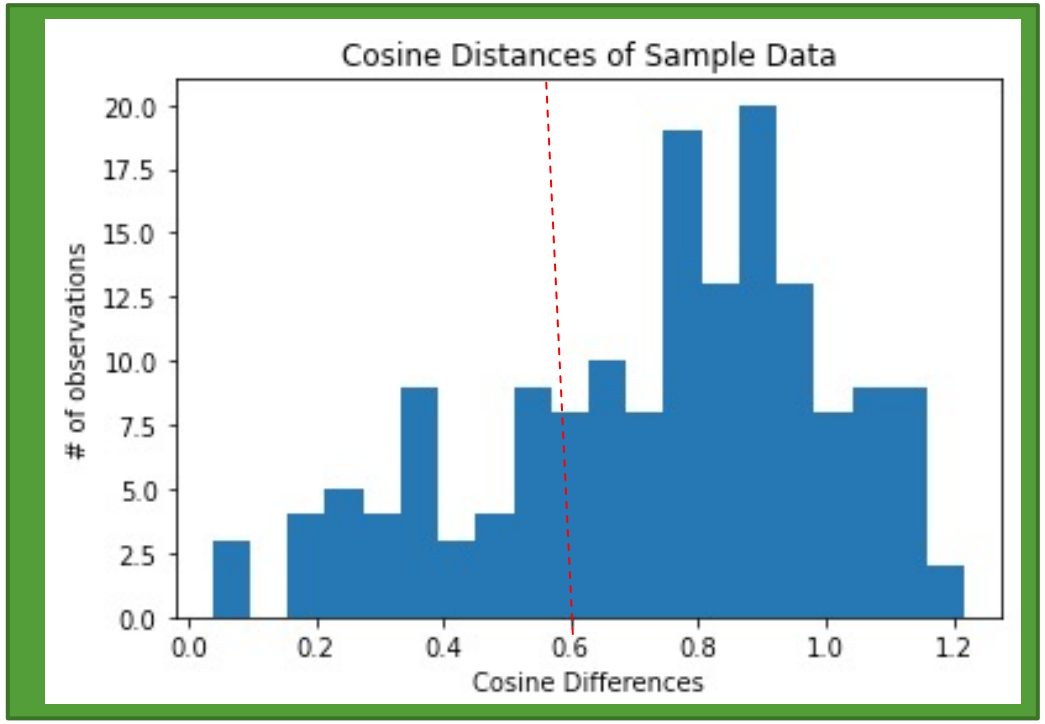
Test Model on New Data

Small sample set of new data (2021 sample of 160 Q & A's) random manual data samples across Amazon food categories

Testing the Model

1) 160 data points
2) Applied to previous corpus and dictionary
3) Results similar (76% correct previous data, 72% new data)
4) Consistent results of Cosine, levenshtein

```
If correctness of answer is based on cosine distance less than:  0.6
Potential wrong answer count:  45
Potential right answer count:  115
Potential wrong answer percentage:  28.125
```



Interactive review of questions and associated answers determined model limit of 5 or lower best predicted actual answers to questions

Final Model Recommendations

- 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) **Amazon can prevent 2/3 of questions from being asked** initially based on Topic Modeling:
    1) 1/3 of questions tied to understanding product contents, counts, or packaging. Make sure allergy potential and gluten contents is clearly identified for customer. Ensure ingredients listing clearly shown and counts or quantities verified
    2) 2/9 of questions on shipping conditions or sourcing. Make sure it is clear to customer origin of shipment (country) and conditions requirements (refrigeration)
    3) 1/9 of questions on match to equipment. Make sure alignment to equipment is clearly stated (pods for coffee makers, etc.)

Of the remaining questions that will always be asked:

2) 1/9 of questions tied to product understanding. Accuracy can be increased based on length and contents. Ensure answer does not contain jargon and is within a window for word count (3-100 words). Add check for entry size on answer box
3) 2/9 of questions are customer complaints, and should be treated separately than simple answers. Utilize modeling to collect these topics for improved customer appreciation

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