M2 Statistics & Econometrics

Final year project

Evaluating Consumer Health Concerns in Online Parenting Communities

Sherman Aline
IRIT, INRAE, SAFFI

April 12 - September 12

Acknowledgements
Thank you to:
Gilles Hubert
Rallous Thomopoulous
Yoann Pitarch
for your guidance and encouragement.

Contents

Abstract

This project was done as a companion project to a study on consumer beliefs about health hazards in infant foods. Rather than surveying consumers, the analysis relied on data gathered from the web, focusing on a forum for mothers in the UK, Netmums.com.

After selecting a subset of posts and classifying them accoridng to product discussed and health hazard discussed, three types of analyses were done.

From Pearson correlation it was found that jarred food is correlated with pesticides and microbes, baby food is correlated with preservatives, and baby formula is correlated with bacteria. It also indicated a need to further filter out Off-topic discussion, particularly of BPA-free plastic bottles, and medical-related issues.

A Paired T-Test and OLS Regression were performed on sentiment measures generated from the texts. Both models had significant coefficients for 'related terms', 'campylobacter', and 'infant formula'. These indicate that infant formula has apositive sentiment, 'related terms' are discussed with objective language, 'campylobacter' has negative sentiment. The T-Tests also found significance across all four measures for 'bisphenol a', inidicating positive sentiment, subjective language, and confident tone. In the OLS model, 'related terms' was confident across all four measures, with contradictory sentiments, unconfident tone, and objective language.

Future work involves investigating sources of variance across measures, better identifying biases through identifying off-topic posts, and considering the relevance of data in the extra category 'related terms'.

0.1 Introduction

I have been working on the SAFFI Project, supported by INRAE and in collaboration with IRIT. The goals of SAFFI (Safe Foods for Infants in the EU and China) is to ensure food safety for infant foods. [1] The project has previously collected survey data on consumer concerns regarding different hazards in different baby food products. The collaboration with IRIT (Institut de Recherche en Informatique de Toulouse) allowed SAFFI to begin an investigation of online sources which could be used to conduct a similar study.

The topics of interest can be separated into two categories: products and hazards. The main interest is to understand which hazards are a concern for consumers, for which products. It's also useful to understand which concerns consumers are knowledgable about, aware of, or more or less worried about. In order to answer these questions the correlation between term occurences for words corresponding to both categories is considered. In order to understand the relationship consumers have with these concers, sentiment measures are generated and we investigate the relationship between these measures and the topics of interest.

Outline Before developing the details of the analyses, a cursory explanation of data collection and feature generation is done. This is in order to give additional emphasis to the goals of the project and clarify sources of bias and the limitations they add in the analysis step. There is an inherent bias present in the initial data collection method, but it's necessary in order to obtain a large enough sample of relevant data. Feature generation consists of sentiment measures from two packages which indicate the emotional sentiment (positive or negative), the modality (confident or unconfident) and subjectivity.

Next, correlation is examined using Pearson correlation coefficients. We also check correlations with a set of words indicating off-topic subjects.

Then, paired T-Tests are performed. We use the fact that forum posts are grouped in threads to create pairs of posts by thread. This test uses topic categories.

Finally, an OLS regression is performed, using raw word-counts, to check the relationship between increased use of words in a post. This method was used alongside T-Tests because both methods have their own strengths and weaknesses.

Not Included Along the way I have had the opportunity to explore possible methods which did not work out, are not finished or are otherwise outside the scope of this report. This includes reading papers on semi-supervised modifications to the Latent Dirichelet Allocation algorithm, writing a package for scraping data, and researching a variety of tools used in Natural Language Processing.

Chapter 1

Data Preparation: Collection, Cleaning, Subpopulation Selection, Feature Creation

1.1 Goal

The aim of this project is to understand the different concerns in relation to one-another. Thus, we do not try to calculate statistics on the entire population. We select a subpopulation which is representative of discussion on only the topics we are interested in: our hazards and products.

1.2 Collection

After considering a variety of sources, the internet forum Netmums was chosen for data collection. Netmums is an internet forum primarily made up of mothers or expecting mothers based in the United Kingdom. [8]

To collect data, I developed a python script which iterates through a set of searches, every item in the set $hazards \times products$. This creates a selection bias in our population sample, which we keep in mind during analysis.

Each search query can return at mos 100 results. Each of these results is a single post in a thread, and so there were some duplicate threads at the end. After gathering all the threads, my script iterates through all posts in each thread and collects post date, time, text, and text of posts which were quoted.

With this method, every thread collected should contain at least one mention of one of our target topics. However it was apparent that much of the data was not relevant and a subpopulation had to be selected.

1.3 Cleaning and Subpopulation Selection

In order to further filter to the data which is relevant to our analysis, I examined the data and developed a criteria to extract a relevant and useful subsample.

- Time - thread contains posts from 2016 or later
- Minimum number of occurrences - a hazard *and* a product occur at least once in the thread
- Term-distance - number of words between product and hazard in the thread is below the 95th percentile.

Term distance It is natural to assume that when words are closer together, they are more likely to be related. In Natural Language Processing, this assumption is quite common, for example the word2vec model relies on a continuous bag of words, i.e. the words surrounding a word, to calculate associations between words. [6, 5] This assumption motivates my development of the term-distance metric.

It is defined for a thread as the minimum distance between a hazard term and a product term, from all posts in that thread. For a post as the minimum word-distance between a hazard term and a product term is

$\forall p$ (products), and h (hazards) in a post P, the term-distance d_P is s.t.

$$d_P = min(\{||p, h||^1; p \in P, h \in P\})$$

where $||p, h||^1 := |i_p - i_h|$ and i_w is the position of a word w in a post

We then calculate the term-distance for a thread by taking the minimum over all post term-distances in the thread.

The term distance criterion is important in selecting a subset which is topic-relevant. If hazard and product are not syntactically close together, this indicates that the product and hazard are not discussed in relation to each other. This is especially important when there are long posts which may actually cover a multitude of subjects.

Minimum Occurence The minimum occurence criterion is important because it we are only concerned with information which can be categorized in a hazard topic and a product topic.

Typo Correction Many of the hazard words are long and difficult to spell. In order to better detect occurences, Levenshtein distance was used to find words which are likely to be typos of our target words.

Levenshtein distance is a metric for measuring similarity between two sequences [7] . In our case, sequences of characters which make up a word. We pass over every word and identify the ones which have short Levenshtein distance relative to the word length.

The Levenshtein distance between two strings a, b is the number of edits (insertion,deletion, or substitution) one has to make to string a to so that it is the same as string b.

A common implementation, (with lengths $|a|$ and $|b|$) is given by $\operatorname{lev}(a, b)$ where

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}(\operatorname{tail}(a), b) \\ \operatorname{lev}(a, \operatorname{tail}(b)) \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

where the tail of some string x is a string of all but the first character of x, and $x[n]$ is the n th character of the string x, starting with character 0.

In our use case, words are identified as typos and corrected if $(|word| - lev(word))/|word| <= 0.8$

1.4 Feature Creation

1.4.1 Classification

After selection of our relevant subsample, topic classification is performed. Each post is assigned a topic from two sets: product and hazards. Classification is done based on the most-occuring terms in the post.

This approach is highly accurate, but many posts are classified as NA due to having zero occurences of a product or a hazard. We use this approach to keep accuracy in exploring the data, as using a machine learning method may begin to incorporate other features into the data which are what we are trying to measure. Interfering with our measurements.

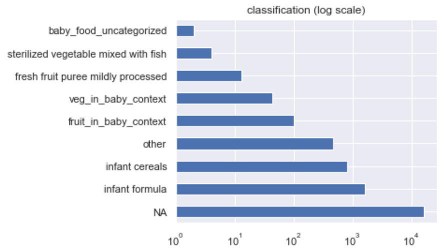
Details Due to small sample sizes in product categories related to baby food, additional categories were created. Often times, specific baby food brand are mentioned which indicates that a baby food is being discussed, but it is not feasible to categorize as vegetable or fruit. This introduces the class baby_food_uncategorized. Additionally, posters may discuss a type of baby food without explicitly stating that it is baby food: it may be obvious by the context of the thread. For these cases we have two categories veg_in_baby_context and fruit_in_baby_context. Both categories refer respectively to occurences where a post mentions a baby in it, or a specific brand of baby food, as well as a mentioning a fruit or vegetable.

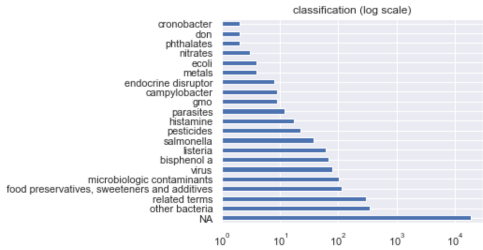
Additionally, an additional placeholder topic category was created in hazards for other possible terms which aren't relevant to current SAFFI hazard interests but could be considered. This topic, 'related terms' counts the following terms:

- carcinogen
- chemicals
- toxic
- toxin
- poisonous
- fungus
- food poisoning
- hazard
- EFSA
- European Food Safety Authority

Possible Improvements In the future a semi-supervised modification to LDA could be used, or a supervised approach which trains on existing labels.

Results





1.4.2 Sentiment

Understanding the Metrics

Two packages are used for sentiment metrics: NLTK Vader and Pattern.[9, 10] Both of these use a lexical approach, which means they have a large dictionary of words with scores. NLTK has been trained on social media data, and is designed specifically for this purpose.[11] Pattern is metrics are trained on books and Wikipedia entries. [13, 14] NLTK sentiment is more relevant for our purpose, but we use Pattern sentiment to compare and also use some other metrics with Pattern offers.

NLTK Vader Sentiment Analysis Vader Sentiment is a multi-step process.

The first step uses a pre-trained classifier to build three different features: negative, neutral, and positive. Each word is classified into on of these categories, and the feature is the proportion of words in the text which were in that class. The sum of these scores should always be 1.

The second step detects grammar patterns and other indicators .1 in the text to weigh higher or lower on different features. With these weighted features, a compound score is created.

- neg $\in [0, 1]$, neu $\in [0, 1]$, pos $\in [0, 1]$
- neg + neu + pos = 1
- f(neg,neu,pos) =compound$\in [-1, 1] - (negative, positive)$

Pattern Measures

Sentiment Pattern's sentiment metric is similar to NLTK, but trained on a dataset of books [13] and it doesn't have the added grammar logic. It has only one class, and takes the average for the sentiment score over all words in the text given.

- Pattern Sentiment $\in [-1, 1]$ - negative/positive

Subjectivity The subjectivity measure is similar to sentiment, relying on hand-tagged scores form a lexicon.

- Subjectivity $\in [0, 1]$, (0 = objective, 1 = subjective)

Modality In the words of the developers, "modality is used to express possibility." In other words, how credulous the author is. The scoring algorithm detects important phrases which indicate the truth or uncertainty, words like "maybe", "allegedly", or "truth". It was trained on Wikipedia articles [14]

- Modality $\in [-1, 1]$, (-1 = unsure, 1 = sure)
- Pattern Modality $\in [-1, 1]$, (-1 = unconfident, 1 = confident)
- Pattern Subjectivity $\in [0, 1]$, (0 = objective, 1 = subjective)

Chapter 2

Analysis

2.1 Summary

Three different methods were used. Co-occurences and correlations were used to see determine which products-hazard pairs are most prevalent in discussion. Paired T-Tests based on topic classification were performed to see which hazards and products have a significant relationship with our sentiment metrics. This indicates the consumer's relationship with the topic, but has sensitive to bias from correlated topics. Lastly, regressions were performed on the term-counts. This can strengthen the indication of a relationship between sentiments and a topic, and identify if other co-occurring off-topic subjects are influencing sentiments.

Specifics of each method are discussed in respective sections.

2.2 Co-Occurences & Correlation

2.2.1 Method

Correlation Correlation is calculated using the Pearson sample correlation coefficient, often just called R. Resulting coefficients are filtered to show only statistically significant results at the 5% level, then further filtered to those with a correlation coefficient with magnitude above 0.1 (or in the last category, 0.08).

The closer the coefficient is to 1, the closer the two variables are to a perfect linear relationship.

- H_0: x and y do not have a linear relationship
- H_1: x and y have a linear relationship

Formula For a correlation coefficient r and two variables x, y

$$r = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

Assumptions and Test Statistic Pearson correlation coefficient relies on the assumption that the two values follow a bivariate normal distribution. [4] Thus for calculating the p-value, the distribution of r is assumed to be

$$f(r) = \frac{(1 - r^2)^{n/2-2}}{\mathrm{B}(\frac{1}{2}, \frac{n}{2} - 1)}$$

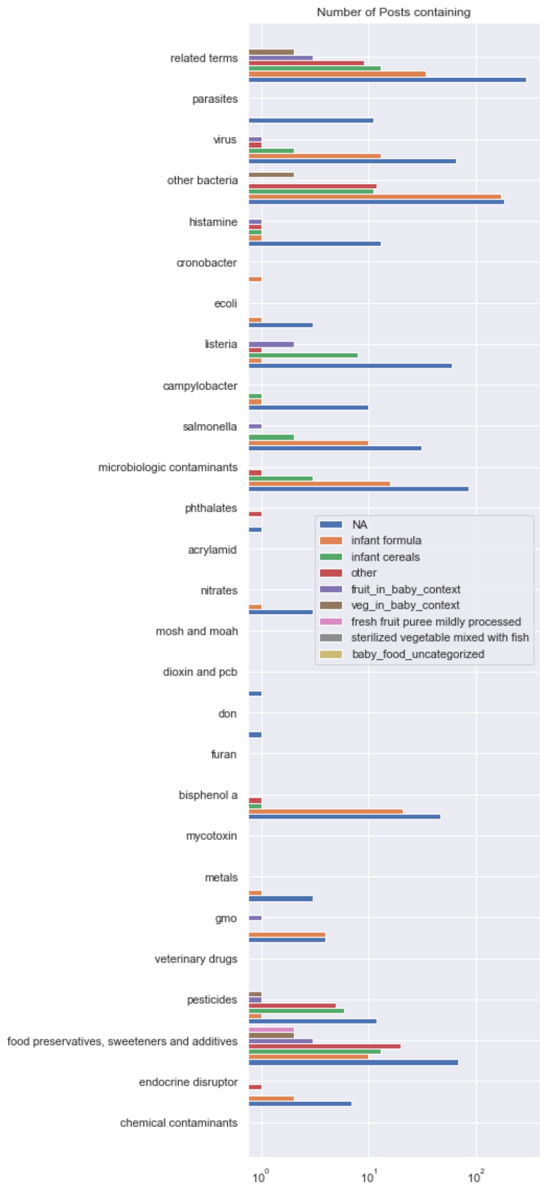
where B is the beta distribution. [12]

2.2.2 Results

Co-Occurence Counts

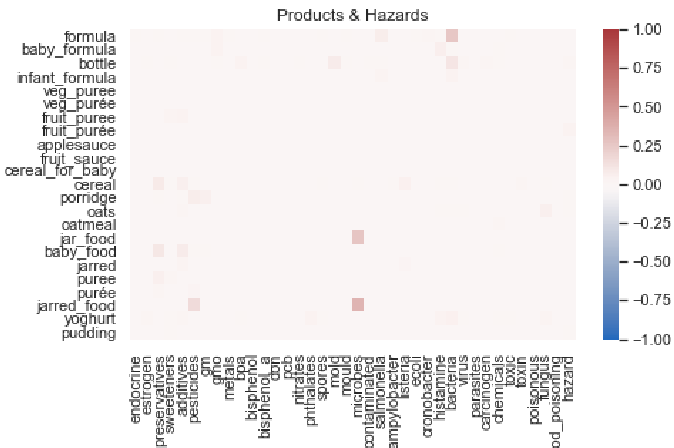
Co-occurence refers to when two words occur in the same space. In our case our space is of an individual post. We use our product-categories to see how hazard are distributed across different product categories.
First, a plot was made of the number of posts which containing words from each hazard-topic with a legend for each product-topic. This allows us to compare if some hazards are more relevant for a specific product while also comparing the hazards among themselves. The x-axis, count of posts, is on log scale which helps us to compare levels in low-count categories.
The NA type is most represented, indicating a large amount of discussion of hazards outside the context of the products we are interested in. Bisphenol A is very prevalent among discussion of infant formula, but not other products. Preservatives, sweeteners and additives are preavalent among all categories.

- x-axis: number of posts containing words related to a specific hazard
- y-axis: hazard (count data - a post can contain more than one hazard term!)
- legend: product (categorical)



Product-Hazard Correlation

- positive correlation: two words occur often in the same post
- negative correlation: two words occur in different posts from each other



bacteria	
bottle	0.117102
formula	0.262494
pesticides	
jarred_food	0.164026
microbes	
jar_food	0.267132
jarred_food	0.353464
preservatives	
baby_food	0.105748

Above is a heatmap of all product-hazard correlations, with a table of those which were significant at the 5% level on the right.

Off-Topic - Hazard Correlation

Three methods of obtaining an optimal set of out-of-topic indicative words were used: Highest-Count, Document-Frequency (DF) Filtered, and Noun-Filtered.

The **goal** is to identify topics which are over-represented in our data for a certain hazard, in order to **identify possible biases**. In particular, medical ailments are of interest, as they come up in the context of allergic reactions of indigestion.

Highest-Count Off-Topic Correlation This approach has no filtering, only taking the most common words.

- free - bpa: high correlation, which may explain some of the positive sentiment seen for ‘bisphenol a’ (BPA)
- bisphenol_a and bpa are both also correlated with bottles. Bottles are used for feeding infant formula and so an indication of concern about BPA in formula is likely to actually be related to plastic baby bottles.

No medical related correlations

Tables

additives	
children	0.117997

bacteria	
bottle	0.117102
bottles	0.125326
formula	0.262494
make	0.120598
water	0.274756
way	0.105163

bpa	
bottles	0.251192
free	0.315302

carcinogen	
products	0.120232

metals	
water	0.228721

parasites	
help	0.142829

toxic	
child	0.132635
food	0.102689
night	0.126593

preservatives	
baby_food	0.105748

bisphenol_a	
bottles	0.128749

DF-Filtered Highest-Count Off-Topic Correlation This approach filters out words which occur in a relatively-large number of documents, and then select the most-occuring from that set.

- listeria is correlated with pregnant, indicating this concern may be primarily in the context of pregant mothers.
- sweeteners - sleep are correlated. Maybe because sugar keeps some children awake?

Medical Related Correlations:

- allergy - histamine
- reflux - histamine

Tables

additives	
bran	0.117717
sleep	0.190053

bacteria	
add	0.128479
bacteria	1.000000
boiled	0.172755
fridge	0.137052
hot	0.167388

histamine	
allergy	0.170797
reflux	0.269144

listeria	
pregnant	0.115364

metals	
body	0.139542

nitrates	
sugar	0.107818

parasites	
body	0.169982
eggs	0.534119
helps	0.153506
makes	0.108331

preservatives	
baby_food	0.105748

salmonella	
bacteria	0.109417
eggs	0.114963
risk	0.108276

sweeteners	
sleep	0.100503
sugar	0.119116

toxic	
sleep	0.225466

Noun-Filtered, DF-Filtered, Highest-Count Off-Topic Correlation This approach detects part of speech and filters out words which are not nouns, along with filtering by document frequency. Because this approach seemed to give the lowest number of low-information words which don't indicate a topic, the correlation threshhold was also lowered.

- Words with correlation above 0.08
- All BPA words are related to bottles: glass, or brand names
- carcinogen - amazon related to online shopping on Amazon.com?

Medical Related Correlations:

- toxic - development the development of the fetus? or the young child maybe?
- teeth - toxic
- finger - hazard (possibly in the context of toys hurting fingers? hazard can be health hazard or physical hazard)
- behind - parasites
- bottom - parasites
- infection - parasites
- pains - parasites
- pp - parasites (may refer to penis?)
- antibiotics - bacteria
- eye - bacteria
- infection - bacteria
- poo - bacteria
- calcium - metals
- cancer - metals
- disease - metals
- teeth - metals

- lactose - histamine
- intolerance - histamine
- acid - histamine
- pump - histamine (acid & pump may be in reference to gastric reflux problems?)
- calcium - nitrates
- teeth - nitrates
- flu - virus
- vaccine - virus

Tables

additives	
attention	0.154567
levels	0.111244
light	0.110201

bacteria	
advance	0.101744
antibiotics	0.131642
degrees	0.179950
filter	0.161717
infection	0.124731
poo	0.100007
shot	0.145204
sterile	0.186941
sterilise	0.125454
temp	0.129280

bpa	
glass	0.166784
stock	0.200236
tippee	0.211833
tommee	0.181109

campylobacter	
bug	0.143547
cats	0.113293

carcinogen	
amazon	0.115253

contaminated	
sterile	0.203767

fungus	
sister	0.130914
thrush	0.110206

gm	
methods	0.133862

histamine	
acid	0.287638
answers	0.107889
drugs	0.132111
hug	0.137355
intolerance	0.203755
prescribe	0.103786
prevent	0.135626
pump	0.167789
system	0.118629
treatment	0.137433

listeria	
bug	0.106264
pate	0.135992
serve	0.235575

metals	
business	0.121014
calcium	0.127285
cancer	0.100817
causes	0.124555
disease	0.132759
fight	0.117654
filters	0.112846
flakes	0.138579
level	0.104644
levels	0.138219
teeth	0.336528

microbes	
date	0.265450
jar	0.220664

mould	
bathroom	0.139909
filter	0.106395
filters	0.154279

nitrates	
teeth	0.192429

parasites	
adults	0.102994
bottom	0.385459
garlic	0.298134
growth	0.152018
hands	0.104574
infection	0.257245
medication	0.216130
methods	0.118349
starts	0.175316
state	0.109689
tablets	0.117839
treatment	0.298737
turn	0.147217
worms	0.369533

pcb	
adults	0.108237
perhaps	0.148155

phthalates	
bags	0.208686
flakes	0.164423
freezer	0.172690
pots	0.137278
toys	0.133398
vinegar	0.135321

poisonous	
changes	0.113630
example	0.122095
poison	0.123237
pure	0.118382

salmonella	
runny	0.120097
supermarket	0.102355

spores	
air	0.160614
cats	0.108665
honey	0.133258

sweeteners	
squash	0.111201

toxic	
adult	0.124720
cry	0.181510
development	0.108119
increase	0.106760
relationship	0.151556
response	0.162152
via	0.101725

toxin	
evidence	0.133723
show	0.171064
third	0.251957

virus	
catch	0.109941
effects	0.100590
flu	0.177605
vaccine	0.292481

2.3 T-Test and Box-Plots

2.3.1 Methods

Paired T-Test was ideal in this situation because it does not require independence. We have selection bias at two steps: the step of data collection and the step of sub-sample selection. Thus, we cannot assume independence and more data would need to be collected for an independent T-Test.

A pair is a set of two measurements from two different treatments, with other conditions more or less constant. Best described by Larsen & Marx, "Paired data, then, consist of measurements taken on Treatment X and Treatment Y within each of b pairs. In effect, the paired t test pools the treatment response differences within each pair from pair to pair." [2]

Assumptions This test assumes that the distribution of the difference $d_i = x_i - y_i$ is Normal. This was verified by visually checking the histograms of the sample distributions.

Construction of Pairs In order to construct a sample for a paired T-Test, the posts are collected by thread. Within each thread, the mean for NA-classified posts and the mean for hazard-classified posts is taken. So our population for each T-Test is limited by the number of threads which contained at least one post with class [topic] and one post with class NA.

Sample Size by Hazard Category			
related terms	167	parasites	8
other bacteria	161	gmo	7
food preservatives, sweeteners and additives	64	endocrine disruptor	7
microbiologic contaminants	48	campylobacter	6
virus	44	bisphenol a	6
listeria	33	metals	3
salmonella	23	ecoli	3
pesticides	16	nitrates	2
histamine	13	phthalates	1

Sample Size by Product Category			
infant formula	268	veg__in__baby__context	27
infant cereals	167	fresh fruit puree mildly processed	11
other	109	sterilized vegetable mixed with fish	2
fruit__in__baby__context	55	baby__food__uncategorized	1

Hypotheses

- H_0: $\mu_d = 0$
  - (sentiment in NA posts and sentiment in target topic posts are the same)
- H_1: $\mu_d \neq 0$
  - (sentiment in NA posts and sentiment in target topic posts are different)

Test Statistic

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{b}}$$

where t follows a Student t-distribution with $b-1$ degrees of freedom.

Box plots of quantiles are also shown as a visual aid to show the relationship between the baseline (NA) and topic sentiments. The vertical red line is the median for the set of all NA posts.

Possible Improvements There are also weaknesses to this approach. If a thread's main topic is a hazard, it is more likely that an NA-classified post just didn't mention the hazard by name. The mean of NA posts in the thread will be biased in the direction of the hazard-classified posts and the relationship will be *underestimated*. If a hazard is mentioned once, off-topic, in a thread of a completely different topic then the significance of the result may be *overestimated*. Samples sizes are also smaller than with other methods.

2.3.2 Results

Summary + and − indicate above or below the baseline (NA), significant is at the 5% level

Pattern Sentiment Significant Results

	Hazards	Products
+	related terms , other bacteria, food preservatives, campylobacter	infant formula, infant cereals , veg in baby context
−		

NLTK Sentiment Significant Results

	Hazards	Products
+		sterilized vegetable mixed with fish
−	related terms, other bacteria, listeria, campylobacter	infant formula

Modality Significant Results

	Hazards	Products
+		
−	related terms, other bacteria, pesticides, parasites	other , fruit in baby context

Subjectivity Significant Results

	Hazards	Products
+		
−	related terms, listeria	

significant at 5% level

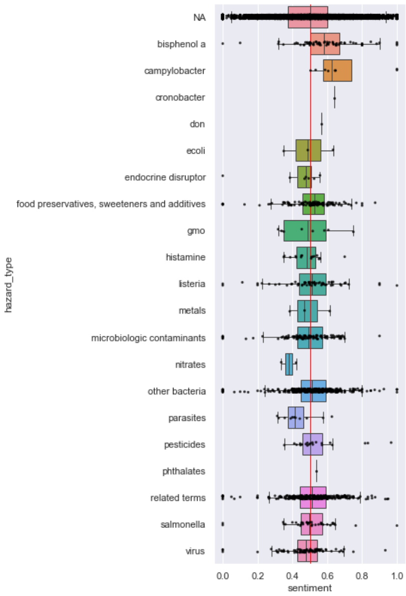
Interestingly, both sentiment metrics have opposite results for hazards. Pattern Sentiment classifies many terms above baseline, and NLTK classifies as below baseline. 'Related terms' is significant across all four metrics, unsurprising as it has a large sample size. Listeria has sentiment below baseline for NLTK metric.

Related terms, other bacteria pesticides, parasites, other, and fruits mentioned in context of baby food are all above baseline modality. This means that posts mentioning these sound *less confident* than those without. No topics have relatively positive modality.

Related terms and listeria have subjectivity below baseline. This means posts related to these topics are *more* objective, that is they seem to be sharing factual information rather than opinions.

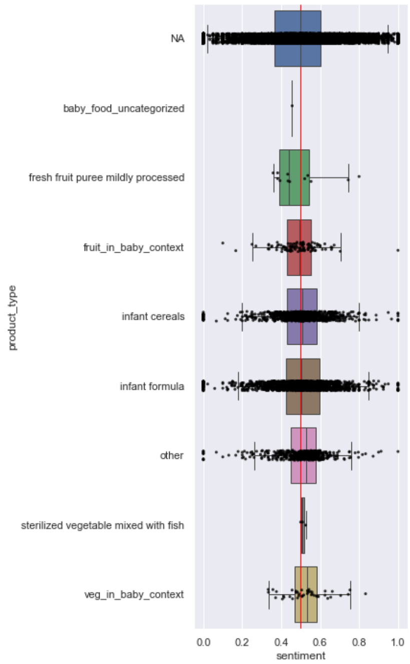
Pattern Sentiment

Hazard



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
related terms	4.037213	0.000082	True
other bacteria	3.927911	0.000127	True
food preservatives, sweeteners and additives	2.007978	0.048938	True
microbiologic contaminants	0.404103	0.687970	False
virus	0.189491	0.850600	False
listeria	1.158046	0.255416	False
salmonella	0.338822	0.737956	False
pesticides	1.604465	0.129455	False
histamine	0.201512	0.843673	False
parasites	0.258988	0.803095	False
gmo	0.375713	0.720055	False
endocrine disruptor	-0.793897	0.457495	False
campylobacter	4.208050	0.008424	True
bisphenol a	0.440058	0.678267	False
metals	0.346728	0.761879	False
ecoli	-0.097355	0.931322	False
nitrates	-2.345835	0.256533	False
phthalates	nan	nan	False
don	nan	nan	False
cronobacter	nan	nan	False

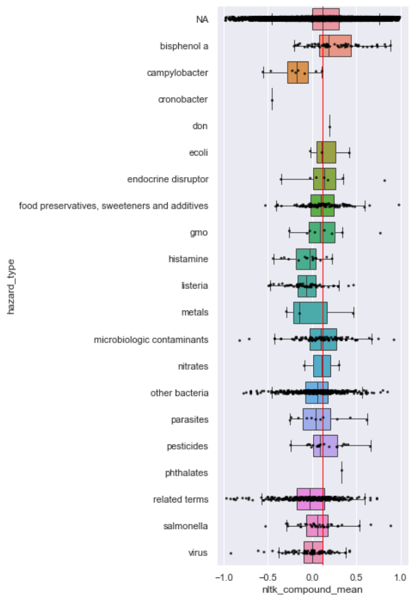
Product



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
infant formula	3.615874	0.000358	True
infant cereals	3.594027	0.000429	True
other	2.301356	0.023291	True
fruit_in_baby_context	1.170302	0.247018	False
veg_in_baby_context	3.961733	0.000517	True
fresh fruit puree mildly processed	0.933480	0.372563	False
sterilized vegetable mixed with fish	1.415220	0.391613	False
baby_food_uncategorized	nan	nan	False

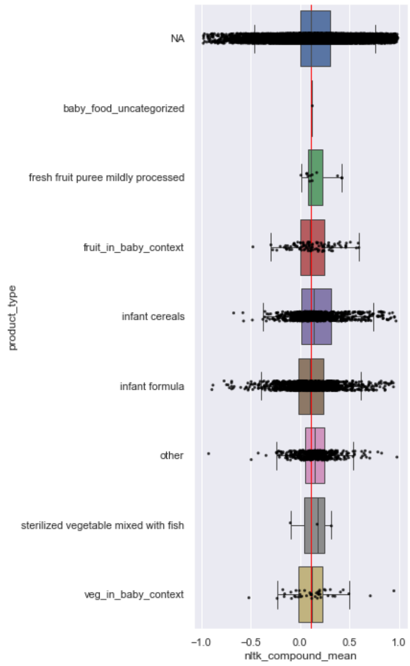
NLTK Sentiment

Hazard



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
related terms	-6.728714	0.000000	True
other bacteria	-4.290691	0.000031	True
food preservatives, sweeteners and additives	-1.781534	0.079645	False
microbiologic contaminants	-1.838208	0.072356	False
virus	-1.532879	0.132631	False
listeria	-5.059567	0.000017	True
salmonella	-1.539032	0.138057	False
pesticides	0.321818	0.752032	False
histamine	-2.105863	0.056949	False
parasites	0.400044	0.701051	False
gmo	-0.076146	0.941779	False
endocrine disruptor	0.590700	0.576279	False
campylobacter	-3.117030	0.026341	True
bisphenol a	1.146081	0.303633	False
metals	-0.309820	0.785999	False
ecoli	0.794296	0.510300	False
nitrates	0.018269	0.988371	False
phthalates	nan	nan	False
don	nan	nan	False
cronobacter	nan	nan	False

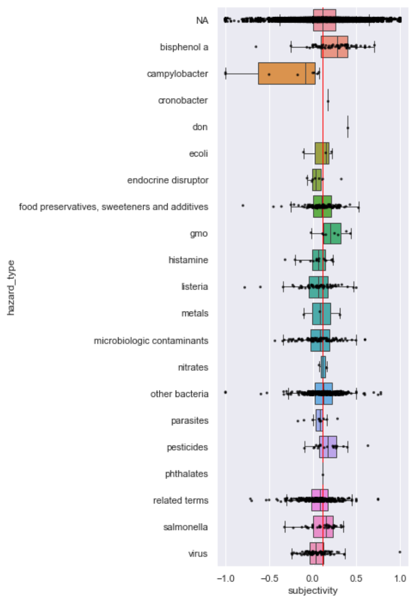
Product



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
infant formula	-2.482138	0.013675	True
infant cereals	1.825320	0.069750	False
other	1.562551	0.121084	False
fruit_in_baby_context	-1.566952	0.122968	False
veg_in_baby_context	0.028637	0.977373	False
fresh fruit puree mildly processed	0.121083	0.906024	False
sterilized vegetable mixed with fish	23.433411	0.027151	True
baby_food_uncategorized	nan	nan	False

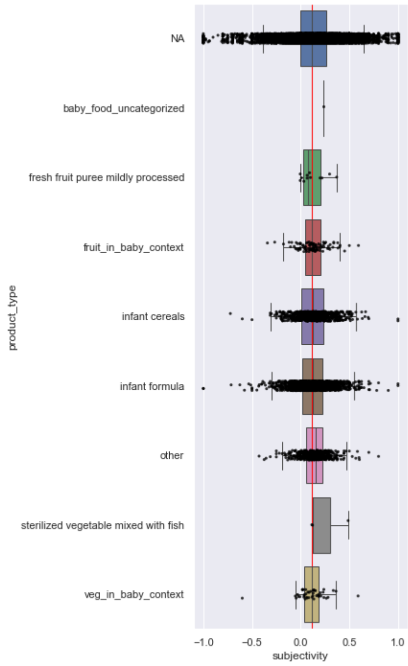
Subjectivity

Hazard



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
related terms	-2.616599	0.009701	True
other bacteria	-0.934891	0.351253	False
food preservatives, sweeteners and additives	-1.697217	0.094591	False
microbiologic contaminants	-1.431580	0.158881	False
virus	-0.416127	0.679386	False
listeria	-2.881635	0.007010	True
salmonella	-1.011015	0.323000	False
pesticides	1.209050	0.245357	False
histamine	-0.252553	0.804888	False
parasites	-0.489177	0.639671	False
gmo	1.159587	0.290273	False
endocrine disruptor	-0.666415	0.529921	False
campylobacter	-1.604677	0.169471	False
bisphenol a	0.448035	0.672874	False
metals	-0.100328	0.929235	False
ecoli	-0.399084	0.728412	False
nitrates	-0.232418	0.854619	False
phthalates	nan	nan	False
don	nan	nan	False
cronobacter	nan	nan	False

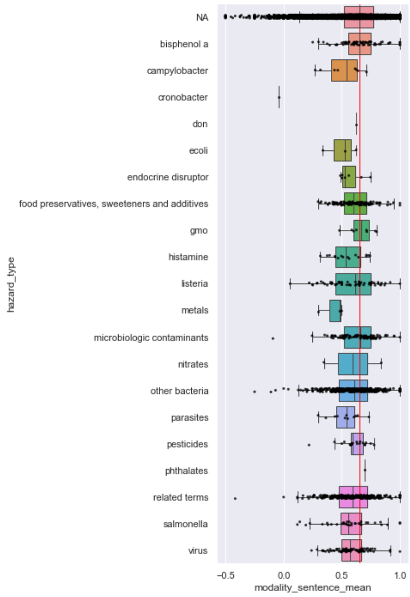
Product



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
infant formula	-0.739547	0.460224	False
infant cereals	1.056269	0.292380	False
other	0.069553	0.944678	False
fruit_in_baby_context	-0.431051	0.668147	False
veg_in_baby_context	0.437524	0.665344	False
fresh fruit puree mildly processed	-0.614009	0.552921	False
sterilized vegetable mixed with fish	1.335795	0.409102	False
baby_food_uncategorized	nan	nan	False

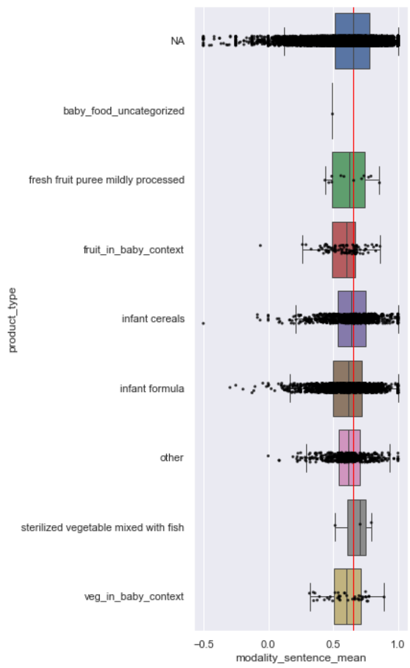
Modality

Hazard



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
related terms	-3.788898	0.000211	True
other bacteria	-2.374733	0.018747	True
food preservatives, sweeteners and additives	-0.652366	0.516539	False
microbiologic contaminants	-0.893776	0.375996	False
virus	-1.029829	0.308848	False
listeria	-0.811809	0.422899	False
salmonella	-0.420211	0.678407	False
pesticides	-2.321150	0.034772	True
histamine	-1.254457	0.233560	False
parasites	-3.628813	0.008411	True
gmo	0.059927	0.954160	False
endocrine disruptor	-0.083102	0.936474	False
campylobacter	-1.661147	0.157575	False
bisphenol a	-0.188342	0.858014	False
metals	-2.830683	0.105430	False
ecoli	-1.139864	0.372458	False
nitrates	-0.100979	0.935932	False
phthalates	nan	nan	False
don	nan	nan	False
cronobacter	nan	nan	False

Product



	T-Statistic	p-value (two-sided)	Reject H0 (two sided) at alpha 0.05
infant formula	-1.502345	0.134190	False
infant cereals	-0.980294	0.328367	False
other	-2.848174	0.005265	True
fruit_in_baby_context	-3.213546	0.002213	True
veg_in_baby_context	-0.912022	0.370136	False
fresh fruit puree mildly processed	-0.619049	0.549726	False
sterilized vegetable mixed with fish	0.121015	0.923332	False
baby_food_uncategorized	nan	nan	False

2.4 Regression

2.4.1 Method

The specification for the regression is OLS with some control variables. The dependent variable is one of the sentiment metrics, with independent variables as term-counts for each topic and a set of control variables from the term-counts for nouns. NA-classified observations are dropped because they contain unobservable correlated variables, because the topics contained in them are unknown.

$$Y = X + Z$$

where Y is the $n \times 1$ vector of sentiment measures, X is $n \times m$ matrix of term-counts for all relevant topics and Z is the $n \times k$ matrix term-counts for control topics.

This specification was chosen for its tractibility and ease of interpretation. Other specifications were considered but each had it's own limitations. .2

Control topics Z were selected using the noun and count filtering method mentioned earlier. This allows us to find topics which might be biasing our prediction. Our results show that the words which correlate with our metrics are often negatively or positively charged word, which are likely to be directly used in the sentiment estimation algorithm.

Practical Difference with T-Test This regression relies on numeric data: count data for word occurences. Thus, coefficients are calculate for the linear relationship between repeated use of a word in a post and the sentiment of that post. This differs from the T-Test done earlier, which relies on topic-categories and test for the difference between posts which are in a category vs. not in that category. Because our categories were classified based on word-counts, this is almost like switching the word counts to a dummy variable for having any occurences or no occurences.

Possible Improvements This specification could be improved by choosing a different method for identifying other topics. Latent Dirichlet Allocation [3] is often used for unsupervised classification and would be useful for our task. However, it is necessary to test whether the unsupervised classification correlates to heavily with our existing hazard and product categories. There exist semi-supervised modifications to LDA which could help ensure this doesn't happen.

2.4.2 Results

Summary - Main Topics + and − indicate above or below the baseline (NA), significant is at the 5% level

Pattern Sentiment Significant Results

	Hazards	Products
+	bisphenol a, campylobacter	infant formula
−	endocrine disruptor	

NLTK Sentiment Significant Results

	Hazards	Products
+	bisphenol a	
−	related terms, listeria, campylobacter	infant cereals

Modality Significant Results

	Hazards	Products
+	bisphenol a	
−	cronobacter	

Subjectivity Significant Results

	Hazards	Products
+	bisphenol a	
−	related terms, campylobacter	

significant at 5% level

With **NLTK sentiment** we find that listeria, virus, and cronobacter have significant negative relationships to sentiment. That is, the more that therms related to these topics are used, the more negative a post is.

Pattern sentiment agrees with this result for bisphenol a, but contradicts it for campylobacter. It also identifies negative sentiment for endocrine disruptor.

In **modality** bisphenol a is found to have a positive coefficient and cronobacter a negative coefficient. This means the more bisphenol a is mentioned, the more likely a post is to be positive. The opposite for cronobacter. We suspect our results in regards to bisphenol are due to most discussion of bisphenol being in the context of BPA-free plastics. Refer to the section "Highest-Count Off-Topic Correlation" on correlations, where 'free' is correlated with bisphenol.

In **subjectivity**, bisphenola a also has a positive coefficient. Related terms and campylobacter have negative coefficients. This means that the more that BPA is mentioned in a post, the *more* subjective the post is. The opposite is true for related terms and campylobacter.

NLTK and Pattern both have and positive coefficient which is significant at the 5% level. NTLK has significant results with a negative coefficient for campylobacter, listeria, and related terms.

Summary - Control Terms (Off Topic) We find several common words in the control set which are positively or negatively charged. Most apprent are "wise", "funny", and "crap". These words should be eliminated from further tests as they are likely to be directly used by the NLTK and Pattern sentiment algorithms.

Pattern Sentiment

Hazard

Dep. Variable:	sentiment	**R-squared:**	0.287
No. Observations:	1180	**Adj. R-squared:**	-0.263
Df Residuals:	666	**F-statistic:**	0.5221
Df Model:	513	**Prob (F-statistic):**	1.00
Covariance Type:	nonrobust	**Log-Likelihood:**	697.95

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.479200	0.015	32.892	0.000000	0.451	0.508
chemical contaminants	-0.000000	5.43e-15	-0.147	0.883000	-1.15e-14	9.87e-15
endocrine disruptor	-0.197100	0.095	-2.085	0.037000	-0.383	-0.011
food preservatives, sweeteners and additives	-0.009300	0.023	-0.402	0.688000	-0.055	0.036
pesticides	0.062900	0.046	1.375	0.170000	-0.027	0.153
veterinary drugs	0.000000	4.5e-15	0.152	0.879000	-8.14e-15	9.51e-15
gmo	-0.073000	0.084	-0.874	0.383000	-0.237	0.091
metals	0.182400	0.305	0.598	0.550000	-0.416	0.781
mycotoxin	0.000000	4.74e-15	0.026	0.979000	-9.18e-15	9.43e-15
bisphenol a	0.041600	0.017	2.419	0.016000	0.008	0.075
furan	-0.000000	2.04e-15	-0.175	0.861000	-4.36e-15	3.64e-15
don	0.089600	0.185	0.484	0.628000	-0.274	0.453
dioxin and pcb	0.030200	0.159	0.189	0.850000	-0.283	0.343
mosh and moah	0.000000	5.02e-15	0.231	0.818000	-8.7e-15	1.1e-14
nitrates	-0.106500	0.140	-0.761	0.447000	-0.381	0.168
acrylamid	-0.000000	3.15e-15	-0.506	0.613000	-7.79e-15	4.59e-15
phthalates	-0.070500	0.203	-0.347	0.728000	-0.469	0.328
microbiologic contaminants	-0.006400	0.020	-0.320	0.749000	-0.045	0.033
salmonella	-0.002300	0.032	-0.072	0.942000	-0.066	0.061
campylobacter	0.108600	0.053	2.057	0.040000	0.005	0.212
listeria	0.009200	0.021	0.440	0.660000	-0.032	0.050
ecoli	0.104500	0.130	0.802	0.423000	-0.151	0.360
cronobacter	0.136700	0.187	0.730	0.466000	-0.231	0.504
histamine	0.030500	0.058	0.526	0.599000	-0.083	0.144
other bacteria	0.016200	0.012	1.352	0.177000	-0.007	0.040
virus	-0.040100	0.026	-1.548	0.122000	-0.091	0.011
parasites	0.016200	0.059	0.276	0.783000	-0.099	0.131
related terms	0.026300	0.014	1.830	0.068000	-0.002	0.054

Figure 2.1: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.479200	0.015	32.892	0.000000	0.451	0.508
dad	0.135400	0.049	2.763	0.006000	0.039	0.232
process	-0.194900	0.078	-2.505	0.012000	-0.348	-0.042
wise	0.228600	0.092	2.487	0.013000	0.048	0.409
crap	0.197000	0.084	2.333	0.020000	0.031	0.363
pieces	0.301600	0.137	2.202	0.028000	0.033	0.571
several	-0.205400	0.093	-2.203	0.028000	-0.388	-0.022
name	0.139300	0.065	2.128	0.034000	0.011	0.268
elsewhere	-0.268600	0.131	-2.044	0.041000	-0.527	-0.011
mini	0.303200	0.149	2.039	0.042000	0.011	0.595
mince	-0.241400	0.120	-2.010	0.045000	-0.477	-0.006

Figure 2.2: control terms

Product

Dep. Variable:	sentiment	**R-squared:**	0.121
No. Observations:	3105	**Adj. R-squared:**	-0.050
Df Residuals:	2599	**F-statistic:**	0.7063
Df Model:	505	**Prob (F-statistic):**	1.00
Covariance Type:	nonrobust	**Log-Likelihood:**	1630.9

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.485700	0.006	84.911	0.000000	0.474	0.497
infant formula	0.006400	0.003	2.043	0.041000	0.000	0.013
sterilized vegetable mixed with fish	0.024100	0.073	0.329	0.742000	-0.120	0.168
fresh fruit puree mildly processed	0.013300	0.029	0.454	0.650000	-0.044	0.071
infant cereals	-0.004100	0.005	-0.808	0.419000	-0.014	0.006
other	-0.001600	0.006	-0.281	0.779000	-0.012	0.009

Figure 2.3: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.485700	0.006	84.911	0.000000	0.474	0.497
funny	0.103200	0.032	3.274	0.001000	0.041	0.165
super	0.053000	0.019	2.838	0.005000	0.016	0.090
mini	0.089600	0.033	2.713	0.007000	0.025	0.154
woman	-0.089500	0.035	-2.560	0.011000	-0.158	-0.021
sad	0.111400	0.050	2.237	0.025000	0.014	0.209
option	0.035800	0.016	2.199	0.028000	0.004	0.068
christmas	-0.060200	0.028	-2.153	0.031000	-0.115	-0.005
mummy	0.057100	0.027	2.095	0.036000	0.004	0.111
crap	0.100300	0.049	2.054	0.040000	0.005	0.196
waste	-0.069800	0.034	-2.044	0.041000	-0.137	-0.003

Figure 2.4: control terms

NLTK Sentiment

Hazard

Dep. Variable:	nltk_compound_mean	**R-squared:**	0.414
No. Observations:	1180	**Adj. R-squared:**	-0.037
Df Residuals:	666	**F-statistic:**	0.9171
Df Model:	513	**Prob (F-statistic):**	0.850
Covariance Type:	nonrobust	**Log-Likelihood:**	175.25

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.066900	0.023	2.947	0.003000	0.022	0.111
chemical contaminants	0.000000	8.46e-15	0.951	0.342000	-8.57e-15	2.47e-14
endocrine disruptor	0.120800	0.147	0.821	0.412000	-0.168	0.410
food preservatives, sweeteners and additives	0.039300	0.036	1.084	0.279000	-0.032	0.110
pesticides	0.044800	0.071	0.629	0.530000	-0.095	0.185
veterinary drugs	-0.000000	7e-15	-1.940	0.053000	-2.73e-14	1.63e-16
gmo	0.044100	0.130	0.339	0.735000	-0.211	0.300
metals	0.109300	0.475	0.230	0.818000	-0.823	1.042
mycotoxin	0.000000	7.38e-15	0.602	0.547000	-1e-14	1.89e-14
bisphenol a	0.069700	0.027	2.601	0.009000	0.017	0.122
furan	0.000000	3.17e-15	0.763	0.446000	-3.81e-15	8.65e-15
don	0.262900	0.288	0.912	0.362000	-0.303	0.829
dioxin and pcb	0.382600	0.248	1.542	0.124000	-0.105	0.870
mosh and moah	-0.000000	7.82e-15	-1.580	0.115000	-2.77e-14	3e-15
nitrates	-0.056500	0.218	-0.259	0.795000	-0.485	0.372
acrylamid	-0.000000	4.91e-15	-1.011	0.312000	-1.46e-14	4.68e-15
phthalates	0.312900	0.316	0.990	0.323000	-0.308	0.934
microbiologic contaminants	-0.003800	0.031	-0.123	0.902000	-0.065	0.057
salmonella	0.059900	0.050	1.188	0.235000	-0.039	0.159
campylobacter	-0.168500	0.082	-2.050	0.041000	-0.330	-0.007
listeria	-0.077700	0.033	-2.375	0.018000	-0.142	-0.013
ecoli	0.120300	0.203	0.593	0.553000	-0.278	0.518
cronobacter	-0.430300	0.291	-1.476	0.140000	-1.003	0.142
histamine	-0.074000	0.090	-0.819	0.413000	-0.251	0.103
other bacteria	-0.030800	0.019	-1.654	0.099000	-0.067	0.006
virus	-0.078100	0.040	-1.935	0.053000	-0.157	0.001
parasites	-0.012400	0.091	-0.136	0.892000	-0.192	0.167
related terms	-0.118700	0.022	-5.308	0.000000	-0.163	-0.075

Figure 2.5: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.066900	0.023	2.947	0.003000	0.022	0.111
lb	0.319500	0.108	2.960	0.003000	0.108	0.531
fun	0.345600	0.128	2.704	0.007000	0.095	0.597
safety	0.222700	0.082	2.725	0.007000	0.062	0.383
spoon	-0.475400	0.180	-2.642	0.008000	-0.829	-0.122
isnt	0.333300	0.127	2.616	0.009000	0.083	0.583
mam	0.171900	0.067	2.561	0.011000	0.040	0.304
crap	-0.330800	0.131	-2.516	0.012000	-0.589	-0.073
answers	0.343600	0.149	2.303	0.022000	0.051	0.637
learn	0.275100	0.120	2.299	0.022000	0.040	0.510
freezer	0.324000	0.145	2.239	0.026000	0.040	0.608
manage	0.326500	0.151	2.156	0.031000	0.029	0.624
drops	0.434700	0.206	2.112	0.035000	0.031	0.839
tests	0.150300	0.072	2.090	0.037000	0.009	0.292
sun	0.234500	0.113	2.072	0.039000	0.012	0.457
became	-0.345100	0.172	-2.010	0.045000	-0.682	-0.008
supermarket	0.184900	0.092	2.004	0.046000	0.004	0.366
bum	-0.653000	0.327	-1.997	0.046000	-1.295	-0.011
appreciate	-0.302000	0.152	-1.989	0.047000	-0.600	-0.004

Figure 2.6: control terms

Product

Dep. Variable:	nltk_compound_mean	**R-squared:**	0.232
No. Observations:	3105	**Adj. R-squared:**	0.083
Df Residuals:	2599	**F-statistic:**	1.558
Df Model:	505	**Prob (F-statistic):**	5.82e-12
Covariance Type:	nonrobust	**Log-Likelihood:**	517.89

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.136100	0.008	16.627	0.000000	0.120	0.152
infant formula	-0.005600	0.004	-1.260	0.208000	-0.014	0.003
sterilized vegetable mixed with fish	-0.054200	0.105	-0.516	0.606000	-0.260	0.152
fresh fruit puree mildly processed	0.032000	0.042	0.764	0.445000	-0.050	0.114
infant cereals	0.016000	0.007	2.224	0.026000	0.002	0.030
other	-0.006100	0.008	-0.765	0.444000	-0.022	0.010

Figure 2.7: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.136100	0.008	16.627	0.000000	0.120	0.152
super	0.213500	0.027	7.990	0.000000	0.161	0.266
death	-0.229800	0.069	-3.344	0.001000	-0.365	-0.095
dermexa	0.019600	0.006	3.313	0.001000	0.008	0.031
goodness	0.200300	0.066	3.054	0.002000	0.072	0.329
pots	0.065600	0.023	2.903	0.004000	0.021	0.110
cry	-0.154500	0.055	-2.809	0.005000	-0.262	-0.047
one_m8	-0.759600	0.290	-2.617	0.009000	-1.329	-0.190
upset	-0.092300	0.036	-2.596	0.009000	-0.162	-0.023
pains	-0.196300	0.075	-2.620	0.009000	-0.343	-0.049
business	0.181500	0.073	2.484	0.013000	0.038	0.325
struggle	-0.103100	0.042	-2.457	0.014000	-0.185	-0.021
suffers	-0.167500	0.071	-2.343	0.019000	-0.308	-0.027
slaughter	0.226600	0.097	2.334	0.020000	0.036	0.417
yummy	0.128000	0.056	2.279	0.023000	0.018	0.238
pepper	0.138000	0.061	2.260	0.024000	0.018	0.258
feet	0.152500	0.068	2.237	0.025000	0.019	0.286
hell	-0.147000	0.066	-2.225	0.026000	-0.277	-0.017
fun	0.092100	0.042	2.197	0.028000	0.010	0.174
tasty	0.086100	0.039	2.195	0.028000	0.009	0.163
mummy	0.085000	0.039	2.179	0.029000	0.009	0.162
screaming	-0.065600	0.030	-2.153	0.031000	-0.125	-0.006
flare	-0.119500	0.055	-2.159	0.031000	-0.228	-0.011
serious	-0.127600	0.059	-2.150	0.032000	-0.244	-0.011
increase	0.088300	0.042	2.118	0.034000	0.007	0.170
mention	-0.105700	0.050	-2.098	0.036000	-0.205	-0.007
ff	0.103900	0.051	2.036	0.042000	0.004	0.204
rat	-0.183900	0.091	-2.028	0.043000	-0.362	-0.006
except	-0.081400	0.040	-2.024	0.043000	-0.160	-0.003
flu	-0.132000	0.065	-2.025	0.043000	-0.260	-0.004
healthier	0.101300	0.051	1.994	0.046000	0.002	0.201
growth	0.084000	0.043	1.967	0.049000	0.000	0.168

Figure 2.8: control terms

Subjectivity

Hazard

Dep. Variable:	subjectivity	**R-squared:**	0.358
No. Observations:	1180	**Adj. R-squared:**	-0.137
Df Residuals:	666	**F-statistic:**	0.7224
Df Model:	513	**Prob (F-statistic):**	1.00
Covariance Type:	nonrobust	**Log-Likelihood:**	493.21

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.110000	0.017	6.347	0.000000	0.076	0.144
chemical contaminants	0.000000	6.46e-15	0.476	0.634000	-9.61e-15	1.58e-14
endocrine disruptor	0.020500	0.112	0.182	0.856000	-0.200	0.241
food preservatives, sweeteners and additives	0.003300	0.028	0.118	0.906000	-0.051	0.058
pesticides	0.044300	0.054	0.814	0.416000	-0.063	0.151
veterinary drugs	-0.000000	5.35e-15	-1.176	0.240000	-1.68e-14	4.21e-15
gmo	0.071600	0.099	0.721	0.471000	-0.123	0.267
metals	0.077000	0.363	0.212	0.832000	-0.635	0.789
mycotoxin	0.000000	5.64e-15	0.803	0.423000	-6.55e-15	1.56e-14
bisphenol a	0.055800	0.020	2.729	0.007000	0.016	0.096
furan	0.000000	2.42e-15	0.958	0.338000	-2.44e-15	7.08e-15
don	0.376600	0.220	1.710	0.088000	-0.056	0.809
dioxin and pcb	0.195400	0.190	1.031	0.303000	-0.177	0.568
mosh and moah	-0.000000	5.97e-15	-0.108	0.914000	-1.24e-14	1.11e-14
nitrates	-0.023400	0.167	-0.140	0.888000	-0.350	0.304
acrylamid	-0.000000	3.75e-15	-0.827	0.408000	-1.05e-14	4.26e-15
phthalates	-0.046900	0.241	-0.194	0.846000	-0.521	0.427
microbiologic contaminants	-0.017500	0.024	-0.738	0.461000	-0.064	0.029
salmonella	-0.016400	0.039	-0.426	0.671000	-0.092	0.059
campylobacter	-0.365300	0.063	-5.817	0.000000	-0.489	-0.242
listeria	-0.030900	0.025	-1.236	0.217000	-0.080	0.018
ecoli	0.128200	0.155	0.828	0.408000	-0.176	0.432
cronobacter	0.109200	0.223	0.491	0.624000	-0.328	0.546
histamine	-0.012900	0.069	-0.186	0.852000	-0.148	0.123
other bacteria	0.020900	0.014	1.471	0.142000	-0.007	0.049
virus	-0.040400	0.031	-1.310	0.190000	-0.101	0.020
parasites	0.009100	0.070	0.130	0.897000	-0.128	0.146
related terms	-0.051900	0.017	-3.037	0.002000	-0.085	-0.018

Figure 2.9: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.110000	0.017	6.347	0.000000	0.076	0.144
crap	-0.402100	0.100	-4.004	0.000000	-0.599	-0.205
sun	0.260400	0.086	3.013	0.003000	0.091	0.430
baths	-0.280100	0.098	-2.863	0.004000	-0.472	-0.088
nappies	0.201900	0.074	2.744	0.006000	0.057	0.346
dad	-0.152700	0.058	-2.619	0.009000	-0.267	-0.038
wise	0.269200	0.109	2.462	0.014000	0.054	0.484
manage	0.283600	0.116	2.452	0.014000	0.057	0.511
pump	0.178000	0.073	2.443	0.015000	0.035	0.321
woman	0.371700	0.152	2.447	0.015000	0.073	0.670
sign	0.256000	0.113	2.272	0.023000	0.035	0.477
daughters	-0.207300	0.092	-2.242	0.025000	-0.389	-0.026
hands	-0.093700	0.044	-2.145	0.032000	-0.180	-0.008
mam	0.109100	0.051	2.127	0.034000	0.008	0.210
clare	-0.487400	0.234	-2.086	0.037000	-0.946	-0.029
pressure	-0.516100	0.248	-2.080	0.038000	-1.003	-0.029

Figure 2.10: control terms

Product

Dep. Variable:	subjectivity	**R-squared:**	0.171
No. Observations:	3105	**Adj. R-squared:**	0.010
Df Residuals:	2599	**F-statistic:**	1.064
Df Model:	505	**Prob (F-statistic):**	0.178
Covariance Type:	nonrobust	**Log-Likelihood:**	1294.9

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.128900	0.006	20.220	0.000000	0.116	0.141
infant formula	0.004200	0.003	1.211	0.226000	-0.003	0.011
sterilized vegetable mixed with fish	0.075800	0.082	0.928	0.353000	-0.084	0.236
fresh fruit puree mildly processed	0.010800	0.033	0.329	0.742000	-0.053	0.075
infant cereals	0.003400	0.006	0.612	0.541000	-0.008	0.014
other	-0.001600	0.006	-0.263	0.793000	-0.014	0.011

Figure 2.11: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.128900	0.006	20.220	0.000000	0.116	0.141
super	0.083400	0.021	4.007	0.000000	0.043	0.124
neither	-0.121400	0.039	-3.093	0.002000	-0.198	-0.044
flare	-0.137000	0.043	-3.178	0.002000	-0.222	-0.052
p5210	0.216600	0.078	2.785	0.005000	0.064	0.369
hows	0.165100	0.061	2.685	0.007000	0.045	0.286
sterilise	-0.042500	0.016	-2.674	0.008000	-0.074	-0.011
suffers	-0.142200	0.056	-2.555	0.011000	-0.251	-0.033
beef	-0.102100	0.040	-2.531	0.011000	-0.181	-0.023
mummy	0.075000	0.030	2.469	0.014000	0.015	0.135
reply	0.088500	0.036	2.443	0.015000	0.017	0.160
crazy	-0.106400	0.045	-2.370	0.018000	-0.194	-0.018
shower	0.113500	0.049	2.301	0.021000	0.017	0.210
waking	-0.087600	0.038	-2.314	0.021000	-0.162	-0.013
page	-0.161400	0.070	-2.318	0.021000	-0.298	-0.025
serious	-0.104100	0.046	-2.253	0.024000	-0.195	-0.013
prescription	-0.059000	0.026	-2.243	0.025000	-0.111	-0.007
training	0.094400	0.043	2.170	0.030000	0.009	0.180
partner	0.064200	0.030	2.108	0.035000	0.004	0.124
vinegar	-0.077500	0.037	-2.089	0.037000	-0.150	-0.005
temp	0.029700	0.014	2.092	0.037000	0.002	0.057
shoulder	-0.117200	0.056	-2.090	0.037000	-0.227	-0.007
piece	-0.083100	0.040	-2.056	0.040000	-0.162	-0.004
woman	-0.077100	0.039	-1.978	0.048000	-0.153	-0.001
awake	-0.068500	0.035	-1.967	0.049000	-0.137	-0.000

Figure 2.12: control terms

Modality

Hazard

Dep. Variable:	modality_sentence_mean	**R-squared:**	0.359
No. Observations:	1180	**Adj. R-squared:**	-0.135
Df Residuals:	666	**F-statistic:**	0.7264
Df Model:	513	**Prob (F-statistic):**	1.00
Covariance Type:	nonrobust	**Log-Likelihood:**	620.30

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.613800	0.016	39.449	0.000000	0.583	0.644
chemical contaminants	0.000000	5.8e-15	0.511	0.609000	-8.43e-15	1.44e-14
endocrine disruptor	0.063300	0.101	0.627	0.531000	-0.135	0.262
food preservatives, sweeteners and additives	0.031400	0.025	1.265	0.206000	-0.017	0.080
pesticides	-0.017900	0.049	-0.366	0.714000	-0.114	0.078
veterinary drugs	-0.000000	4.8e-15	-1.084	0.279000	-1.46e-14	4.22e-15
gmo	-0.000400	0.089	-0.005	0.996000	-0.176	0.175
metals	-0.108800	0.326	-0.334	0.739000	-0.748	0.531
mycotoxin	0.000000	5.06e-15	0.649	0.516000	-6.65e-15	1.32e-14
bisphenol a	0.054500	0.018	2.967	0.003000	0.018	0.091
furan	-0.000000	2.18e-15	-0.684	0.494000	-5.76e-15	2.78e-15
don	0.043500	0.198	0.220	0.826000	-0.345	0.432
dioxin and pcb	0.166300	0.170	0.977	0.329000	-0.168	0.501
mosh and moah	-0.000000	5.36e-15	-1.000	0.318000	-1.59e-14	5.17e-15
nitrates	-0.065800	0.150	-0.440	0.660000	-0.359	0.228
acrylamid	-0.000000	3.37e-15	-0.800	0.424000	-9.31e-15	3.92e-15
phthalates	-0.208300	0.217	-0.961	0.337000	-0.634	0.217
microbiologic contaminants	0.025700	0.021	1.207	0.228000	-0.016	0.067
salmonella	0.006000	0.035	0.175	0.861000	-0.062	0.074
campylobacter	0.035100	0.056	0.623	0.534000	-0.076	0.146
listeria	-0.010400	0.022	-0.463	0.644000	-0.054	0.034
ecoli	-0.142100	0.139	-1.022	0.307000	-0.415	0.131
cronobacter	-0.579100	0.200	-2.897	0.004000	-0.972	-0.187
histamine	-0.047800	0.062	-0.771	0.441000	-0.169	0.074
other bacteria	0.001500	0.013	0.116	0.908000	-0.024	0.027
virus	0.036900	0.028	1.332	0.183000	-0.017	0.091
parasites	-0.013400	0.063	-0.214	0.830000	-0.136	0.109
related terms	-0.012300	0.015	-0.803	0.422000	-0.042	0.018

Figure 2.13: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.613800	0.016	39.449	0.000000	0.583	0.644
mince	-0.417200	0.128	-3.253	0.001000	-0.669	-0.165
adults	0.238700	0.105	2.268	0.024000	0.032	0.445
pork	0.807900	0.363	2.224	0.027000	0.095	1.521
disease	-0.133600	0.062	-2.168	0.030000	-0.254	-0.013
texture	-0.282900	0.131	-2.159	0.031000	-0.540	-0.026
happens	0.190200	0.091	2.099	0.036000	0.012	0.368
amazon	0.218200	0.104	2.098	0.036000	0.014	0.422
yummy	0.264000	0.126	2.092	0.037000	0.016	0.512
therefore	0.109100	0.053	2.054	0.040000	0.005	0.213
number	-0.115900	0.057	-2.044	0.041000	-0.227	-0.005
doubt	-0.192600	0.095	-2.036	0.042000	-0.378	-0.007
reaction	-0.120700	0.059	-2.039	0.042000	-0.237	-0.004
n7105	-0.000000	2.33e-15	-2.039	0.042000	-9.32e-15	-1.75e-16
muslims	-0.863100	0.427	-2.021	0.044000	-1.702	-0.025
peppers	-0.859500	0.430	-2.000	0.046000	-1.703	-0.016

Figure 2.14: control terms

Product

Dep. Variable:	modality_sentence_mean	**R-squared:**	0.187
No. Observations:	3105	**Adj. R-squared:**	0.029
Df Residuals:	2599	**F-statistic:**	1.181
Df Model:	505	**Prob (F-statistic):**	0.00654
Covariance Type:	nonrobust	**Log-Likelihood:**	1281.4

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.630100	0.006	98.437	0.000000	0.618	0.643
infant formula	0.000085	0.004	0.024	0.981000	-0.007	0.007
sterilized vegetable mixed with fish	-0.015500	0.082	-0.189	0.850000	-0.176	0.145
fresh fruit puree mildly processed	0.007100	0.033	0.218	0.828000	-0.057	0.071
infant cereals	0.006200	0.006	1.095	0.273000	-0.005	0.017
other	-0.000600	0.006	-0.099	0.921000	-0.013	0.012

Figure 2.15: topic terms

Parameter	coef	std err	t	P>|t|	[0.025	0.975]
const	0.630100	0.006	98.437	0.000000	0.618	0.643
pork	0.078200	0.024	3.301	0.001000	0.032	0.125
freezer	-0.124700	0.036	-3.435	0.001000	-0.196	-0.054
date	-0.060600	0.020	-3.058	0.002000	-0.099	-0.022
almond	-0.102500	0.036	-2.873	0.004000	-0.172	-0.033
advance	-0.058600	0.021	-2.834	0.005000	-0.099	-0.018
afford	-0.123300	0.045	-2.738	0.006000	-0.212	-0.035
fight	-0.139700	0.054	-2.603	0.009000	-0.245	-0.034
anybody	-0.111100	0.045	-2.489	0.013000	-0.199	-0.024
iv	0.092000	0.037	2.470	0.014000	0.019	0.165
struggle	-0.080500	0.033	-2.453	0.014000	-0.145	-0.016
sons	-0.071900	0.029	-2.450	0.014000	-0.130	-0.014
somewhere	-0.085100	0.035	-2.458	0.014000	-0.153	-0.017
spend	-0.080400	0.033	-2.416	0.016000	-0.146	-0.015
feet	0.125500	0.053	2.353	0.019000	0.021	0.230
advise	-0.060800	0.026	-2.339	0.019000	-0.112	-0.010
question	-0.068900	0.030	-2.326	0.020000	-0.127	-0.011
serve	-0.060000	0.027	-2.212	0.027000	-0.113	-0.007
funny	0.077100	0.035	2.187	0.029000	0.008	0.146
hows	0.133500	0.062	2.162	0.031000	0.012	0.255
breastmilk	-0.047600	0.022	-2.127	0.034000	-0.091	-0.004
doubt	-0.085200	0.040	-2.108	0.035000	-0.165	-0.006
berries	0.057400	0.027	2.114	0.035000	0.004	0.111
yummy	0.092200	0.044	2.100	0.036000	0.006	0.178
periods	-0.123400	0.059	-2.098	0.036000	-0.239	-0.008
otherwise	-0.059500	0.028	-2.099	0.036000	-0.115	-0.004
causes	-0.088500	0.043	-2.071	0.038000	-0.172	-0.005
seemed	-0.042800	0.021	-1.993	0.046000	-0.085	-0.001
screaming	-0.047400	0.024	-1.989	0.047000	-0.094	-0.001
mention	-0.078100	0.039	-1.982	0.048000	-0.155	-0.001
area	-0.077900	0.039	-1.978	0.048000	-0.155	-0.001

Figure 2.16: control terms

Chapter 3

Conclusion

Comparing Sentiment Measures In conclusion, we found significant results across all three approaches. In regards to sentiment metrics, NLTK sentiment seemed to be more consistent and so had more results than Pattern's sentiment measure. Interestingly, these two metrics offered contradictory results a number of times, and this is something that should be investigated in the future. For any future work though, it is suggest to use the NLTK metric as it has a larger userbase and is likely to be more reliable.

T-Test For paired T-Tests, Pattern sentiment indicated that 'related terms', 'other bacteria' and 'campylobacter' had relatively positive sentiment, where NLTK indicated negative sentiment for these three. Pattern also indicated 'food preservative' and the product topics 'infant formula', 'infant cereals' and 'veg in baby context' as having relatively positive sentiment. NLTK instead picked up 'sterilized vegetable mixed with fish' as the only product relatively positive sentiment. NLTK also indicated 'infant formula' as having relatively negative sentiment. This is another surprising result, which warrants further investigation.

'Pesticides' and 'parasites' were associated with negative modality, which indicates consumers do not have confident information about them. The same is true for 'fruit in baby context', which represents fruit-based baby foods. Subjectivity reported the smallest number of significant results. 'Related terms' and 'listeria' were indicated as being talked about more objectively than baseline (NA).

Vague Terms are Associated with Uncertainty Also of note, the most vague categories ('related terms', 'other bacteria', 'pesticides', 'parasites') were indicative of relatively low modality, meaning that consumers indicate a lack of confidence when discussing these hazards. This result is logical as people who use less specific words (in contrast to a specific strain of bacteria such as campylbacter) are likely to have weaker knowledge of the topic. Althought these categories are not ideal because they lack specificity, this confirms that their inclusion is useful; less-informed consumers will not be represented when these terms are not considered.

Regression For the Regressions, we found that 'endocrine disruptor' had relatively negative sentiment and 'bisphenol a', 'campylobacter', and 'infant formula' had relatively positive sentiment. NLTK sentiment indicated a similar result for 'bisphenol a', and relatively negative sentiment as mention of 'infant cereals', 'related terms', 'listeria' or 'campylobacter' increase.

Modality and Subjectivity also indicate significant positive coefficients for 'bisphenol a'. This means it is spoken about with confidence and a lot of subjective language. This strengthens the hypothesis that the majority of the data on BPA is in the context of purchasing BPA-free plastic bottles, friendly shopping discussion will be more confident, positive, and subjective.

Modality also indicated that 'cronobacter' is spoken about with less confidence, while the subjectivity measure indicates that 'campylobacter' and 'related terms' is spoken about with objective language.

Implications for Future Data Collection It is noteworthy that campylobacter, cronobacter, and listeria have strong results compared to other topics, despite relatively small samples sizes. This could be related to their method of collection: these are clearly defined topics and are collected solely based on mentions of the topic. Other topics like 'metals', 'virus' and 'food preservatives, sweeteners, and additives' have added complexity in collection due to less well defined topics as well as higher likelihood of capturing off-topic discussion.

Comparing T-Test and Regression The T-Test and Regression results do not perfectly coincide. Both methods have limitations. In regressions, more than one occurence of a hazard is less likely than only one occurence, and so for small samples it may be difficult to fit linearly when there are not many high-count observations. For the Paired T-Test, the sample sizes are limited by the number of threads containing non-classified and successfuly classified posts.

Correlation By checking Pearson correlation, it has also been found that there are some off-topic discussions in our dataset. These need to be investigated further. We also found that jarred food is strongly correlated with microbes, a coefficient of 0.35. Also that formual and bacteria are often mentioned together, with a Pearson correlation coefficient of 0.2.

Future Work Overall, we have found that the largest amount of discussion of hazardous products is involved in baby formula. However in order to better confirm this result, it is necessary to separate out discussion of BPA-free plastics. Additionally, better-defined categories or better methods of capturing topics like viruses is important, but is a unique challenge due to less specific language than terms like 'campylobacter'. This possibly echoes the consumers own lack of knowledge in fields where less specific terms are used, since the less specific categories have a negative modality, indicating lack of confidence.

Bibliography

[1] *About SAFFI*. 2021. URL: https://www.saffi.eu/about-saffi/.

[2] *An Introduction to Mathematical Statistics and its Applications*. Prentice Hall, 2012.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* (2003).

[4] Charles J. Kowalski. "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (1972). eprint: 0035-9254,1467-9876.

[5] Tomas Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546 [cs.CL].

[6] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[7] Gonzalo Navarro. "A guided tour to approximate string matching". In: *ACM computing surveys (CSUR)* (2001).

[8] *Netmums Forum: Pregnancy, Parenting and Family Life Chat*. 2021. URL: https://www.netmums.com/coffeehouse/.

[9] *pattern en Documentation*. 2021. URL: https://github.com/clips/pattern/.

[10] NLTK Project. *Sentiment Analysis*. 2015. URL: https://www.nltk.org/howto/sentiment.html.

[11] NLTK Project. *VADER-Sentiment-Analysis*. 2015. URL: https://github.com/cjhutto/vaderSentiment.

[12] *scipy.stats.pearsonr notes*. 2021. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html.

[13] Tom de Smedt. *Test Corpora*. 2013. URL: https://github.com/clips/pattern/blob/d25511f9ca7ed9356b801d8663b8b5168464e68f/test/corpora/README.txt.

[14] Tom de Smedt and Markus Beuckelmann. *Pattern EN Mood & Modality (Source Code)*. 2017. URL: https://github.com/clips/pattern/blob/master/pattern/text/en/modality.py.

Appendix

.1 Vader Rules

VADER rule-based enhancements include word-order sensitivity for sentiment-laden multi-word phrases, degree modifiers, word-shape amplifiers, punctuation amplifiers, negation polarity switches, or contrastive conjunction sensitivity

.2 Alternative Regression Specifications

Interaction Effects model was tried, but samples are so small that no results are significant. Mixed Effects model was attempted, but small samples caused a problem of collinearity.