Gaining Inference in a Machine Learning NLP Sentiment Analysis:

An Application to the Topic of Genome Editing in Domestic Livestock using Twitter Data

Joseph Navelski *
School of Economic Sciences
Washington State University

October 11, 2022

*** Please do not distribute or cite without permission. ***

Abstract

With the assumption that opinions on social media can give insight into consumer behavior, I use Twitter text data from the United States and machine learning to estimate user sentiment towards genome editing in domestic livestock. With a sample size of 384,452, I find that the average sentiment towards genome editing in domestic livestock is approximately 0.502, on a sentiment index from zero to one, from January 2010 to January 2021. I further analyze user sentiment by the search terms used to procure the data and at the state level, and find the terms closely related to "genome editing," such as "biotechnology," "crispr," and "gene editing," all have similar sentiment levels at 0.503, 0.502, and 0.501. To further motivate prediction results, I bootstrap and formally test if there is a significant different between the stochastic distributions within the search terms and state pairs using a One-Way ANOVA and a Wilcoxon–Mann–Whitney test. Results show that the "dehorning" and "genome editing" factor distributions are significantly different 42.3% of the time with sentiment means of 0.496 and 0.502, respectively, and that Oklahoma and Oregon are significantly different 13.56% of the time with sentiment means of 0.497 and 0.507, respectively. Practitioners should adopt this technique when motivating machine learning prediction results, and policy makers should use these results to gain insights on how a large population in the United States views genome editing in domestic livestock.

Key Words: Big Data, Applied Econometrics and Statistics, Machine Learning, Natural Language Processing (NLP), and Social Media Data

*301H Hulbert Hall, School of Economic Sciences, Washington State University, Pullman, WA 99164-6210, USA. E-mail: joseph.navelski@wsu.edu.

1 Introduction

In this paper, I use social media text data to estimate user sentiment towards genome editing in domestic livestock. With the assumption that opinions on social media can give insight into consumer behavior, I use Twitter data from the United States and machine learning to estimate user sentiment about the term "genome editing" and the terms that surround the topic of genome editing in domestic livestock. This methodology produces a sentiment index from zero to one, and I use this index in a factor analysis to gain insight on how sentiment differs at different factor levels. The two factors I use to analyze user sentiment are the search terms used to procure the data from the Twitter archive and the state in which the tweet came from. Results show that the overall sentiment towards genome editing in domestic livestock is 0.502 with a standard deviation of 0.035, and that the term "genome editing" has a sentiment of 0.501. Results also show that the sentiment between search terms differs with "organic" and "dehorning" having the highest and lowest sentiment ar 0.505 and 0.497, respectively. State sentiment also differ with Oregon having the highest sentiment at 0.503, and with Wisconsin, Alaska and Texas all having the lowest sentiment at 0.498. Another interesting result is that the terms one would assume to be closely related to "genome editing," such as "biotechnology," "crispr," and "gene editing," all have similar positive sentiment level of 0.503, 0.502, and 0.501, respectively.

I also employ a novel methodology to tease out the most probable and significantly different sentiments for each factor pair. Results show that the sentiment between the search terms factor levels and state factor levels are significantly different in each factor set, and that there is a probability of associated with taking inference on the factor levels that are significantly different from each other. Interestingly, "dehorning" and "genome editing" are the terms most probable to have a significant difference between their stochastic distributions with means of 0.496 and 0.502, respectively. Oklahoma and Oregon have the highest probability of having significantly different stochastic distributions with sentiment means of 0.497 and 0.507, respectively. Gaining inference in this application is important because it allows prediction results to be further validated. These results should give insights on how users perceive or accept genome editing in domestic livestock, and policy makers should use these insights when creating and implementing genome editing policies in the domestic livestock arena.

2 Literature Review

There are three recent articles that investigate consumer acceptance and perception towards genome editing in animal food products, or in a more general arena, genetically modified organisms (GMOs). Ortega et al. (2022) study how consumers accept genome edited pork products in China through a geographically dispersed consumer acceptance choice experiment and survey. The study was administered to see if consumers were willing to consume gene-edited and transgenetic pork for the purpose of preventing African swine flu (ASF) outbreaks in the production process. Results show 38% of the respondents were in support of consuming gene-edited pork to prevent ASF, and that 30% of the respondents were in support of transgenetic pork to prevent ASF. These results are obviously important, but the choice experiment was administered through choosing "pictures of food" rather than individuals actually purchasing the food to physically consume, and the study was administered in China and not in the United States. Another related study is that of Tabei

et al. (2020) where the researchers used Twitter data, from 14,066 users, to analyze their sentiment towards genome-edited foods and the labeling policy of Japan's Consumer Affairs Agency. This study concluded that 54.5% to 62.8% of the tweets were negative about the Consumer Affairs Agency's labeling policy towards genome-edited foods. These results are interesting, but they focus more on genome-editing in general and not specifically in animal products. Furthermore, the study is based in Japan, and not the United States.

The most comprehensive study that relates most to this paper's research is that of Wirz et al. (2021). Wirz et al. (2021) utilizes Twitter data to predict the sentiment of users towards GMOs by state, and finds that the sentiments expressed about GMOs is related more towards the topics of the tweets rather than the state level economic indicators like education or political ideology. The researchers used a Twitter dataset with 4,813,197 tweets related to GMOs that were posted from January 1, 2016 to May 1, 2018, and use a nonparametric content analysis software, known as *Crimson Hexagon ForSight*, to predict the sentiment of tweets (Hopkins and King (2010)). The authors find that 41% of the state specific tweets had negative sentiment, 30% were neutral, and 26% were positive, and they present these results for each state in a table. This paper is the gold standard of predicting user's sentiment about GMOs on Twitter, but they do not cover tweets earlier than 2016, do not specifically investigate the difference between the terms that are related to GMOs such as genome editing, and they do not try to take inference between the sentiment levels within each factor they investigate. For this, I try to fill these gaps in the research.

3 Motivation

Measuring consumers perceptions about genome editing in general, let alone in the domestic livestock industry, is difficult when consumer purchasing data on these products is not readily available to the public. Additionally, collecting this data is usually timely and expensive in a time where policy makers face the difficult decision of allowing genome edited animal products into the US market. This paper is intended to inform policy makers about the perception of a large population in the United States that has an opinion on genome edited in the domestic livestock production sector. With the assumption that US Twitter users are consumers with opinions on genome edited foods, I estimate the sentiment of these users to help policy makers make more informed decisions about genome edited foods in general and in the domestic livestock industry.

4 A Sentiment Analysis Using Twitter Data

4.1 Identifying Which Tweets Talk about a Particular Topic

In order to conduct a sentiment analysis using Twitter data, I first identify a set of keywords that relate to the topic of "genome editing in domestic livestock." These key words were identified through literature reviews and through a software called Social Mention. Social Mention is an online software that identifies the key terms that are closely related to a topic people are posting about online. The software scans a large portion of the social media platforms and returns a list of key words. My search results show that **animal welfare, biotechnology, crispr, dairy, dehorning, gene editing, genetically modified, genome editing, genome engineer, GMO,** and **organic**

are the words that are most closely related to the topic of "genome editing in domestic livestock," and I use these words as a basis for analysis.

4.2 Social Media Data

I use an Academic Researcher account with Twitter to procure the social media data used in the analysis. With access to their full archive of public Twitter data, I use their application programming interface (API) to scrape tweets according to the key words listed in Section 4.1. I collect every tweet sent from the US from January 2010 to January 2021 that contains a key word from Section 4.1. This scrape yields a dataset with more than 2 million tweets, but many of the tweets have either a self-identified location that is not available or blank, or the location is specified outside of the United States. Recognizing this, I code a simple text classifier that recognizes and imputes the state abbreviation for each tweet that has a state "name" and/or "abbreviation" clearly listed in the self-identified location of the tweet. This process reduces the dataset to 384,452 observations, and Figure 1 (a) shows the number of tweets over time for each search term, and Figure 1 (b) shows the number of tweets that come from each state over time. One thing to notice in these plots is how about 90.45% of the tweets were acquired by searching for the terms "dairy" and "organic," and that about 18.09% of the tweets come from California. This implies that I need to be careful when taking inference from this sample because the test statistics could be heavily influenced by one portion of the population.[1]

(a) (b)

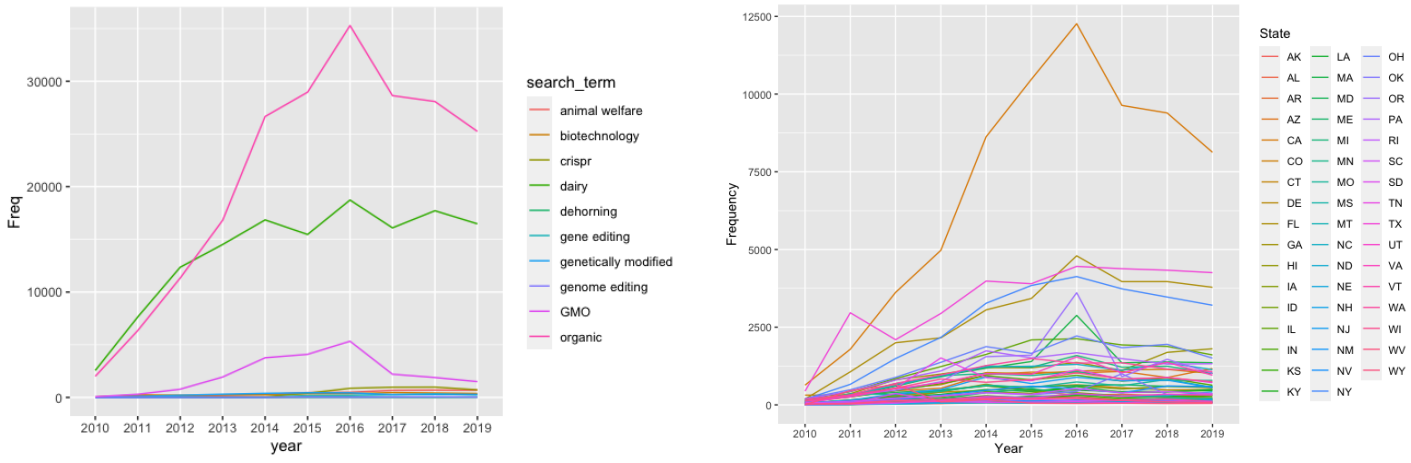

Figure 1: Frequency of Tweets in the United States about Genome Editing

4.3 Statistical Model 1 - Sentiment Prediction

I use a common machine learning classification method, called logistic regression, to predict if a tweet is positive or negative. This classification method uses the text of a tweet to predict its

[1]To address this, I employ bootstrapping techniques that allow me to randomly sample from the entire Twitter "population." This allows me to take inference on smaller sub-samples, and then construct empirical distributions of test statistics to generalize results.

sentiment based on a carefully selected training dataset of positive and negative tweets. I use the industry standard training dataset from Python's Natural Language Tool Kit `NLTK` to train the logistic regression, and the dataset consists of 10,000 tweets, in which 5,000 are positive and 5,000 are negative (Bird et al. (2009)). I train the logistic regression on 8,000 tweets, 4,000 positive and 4,000 negative, and test it on the remaining 2,000 out-of-sample tweets. The trained logistic regression has a classification rate of 98%, and the theoretical statistical model is specified as

$$h(z) = \frac{1}{1 + \exp^{-z}}$$

where $z = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_N x_N$ and the a loss function defined as

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log(h(z(\theta)^{(i)})) + (1 - y^{(i)}) \log(1 - h(z(\theta)^{(i)}))$$

where m is the number of training examples, $y^{(i)}$ is the actual label of the i^{th} training example, and $h(z(\theta)^{(i)})$ is the model's prediction for the i^{th} training example.[2] Each x_i is a separate frequency count of each word in the positive and negative tweets, and all 'stop words" were removed from the analysis. Additionally, all words were "stemmed" using the widely accepted Porter stemming algorithm (or "Porter stemmer"), which is a process for "removing the commoner morphological and inflectional endings from words in English" (van Rijsbergen et al. (1980)). I apply the gradient descent method (Newton-Raphson) of optimization to estimate the governing parameters and then uses these parameter estimates to predict the sentiment of the 384,452 out-of-sample tweets related to genome editing. For simplicity, I define all predicted sentiment as $\hat{y}_i$, instead of $\widehat{h(z)}$, henceforth. This sentient prediction range can be interpreted as a sentiment index with one being the most positive level of sentiment, zero being the most negative, and .5 being the relative indifference level and/or the logistic regressions classification plane.

4.4 Preliminary Results

Overall, the mean sentiment about the search terms related to genome editing is 0.502, which is positive, and the standard deviation is 0.035, which is not very large in magnitude. With that said, the 5th and 95th percentile values are 0.479 and 0.528 respectively, implying that some of the predictions may be outliers or have a slew of many positive or negative words. For analysis validity, I retain these observations, but it should be noted that they could be dropped. Figure 2 is a histogram of the predicted sentiment values for the 384,452 tweets related to genome editing, and in the following sections I do a post-hoc analysis on some of the underlying factors that might influence sentiment.

[2]Note that we can define the loss function for a single observation as $Loss = -1 \times \left(y^{(i)} \log(h(z(\theta)^{(i)})) + (1 - y^{(i)}) \log(1 - h(z(\theta)^{(i)}))\right)$.

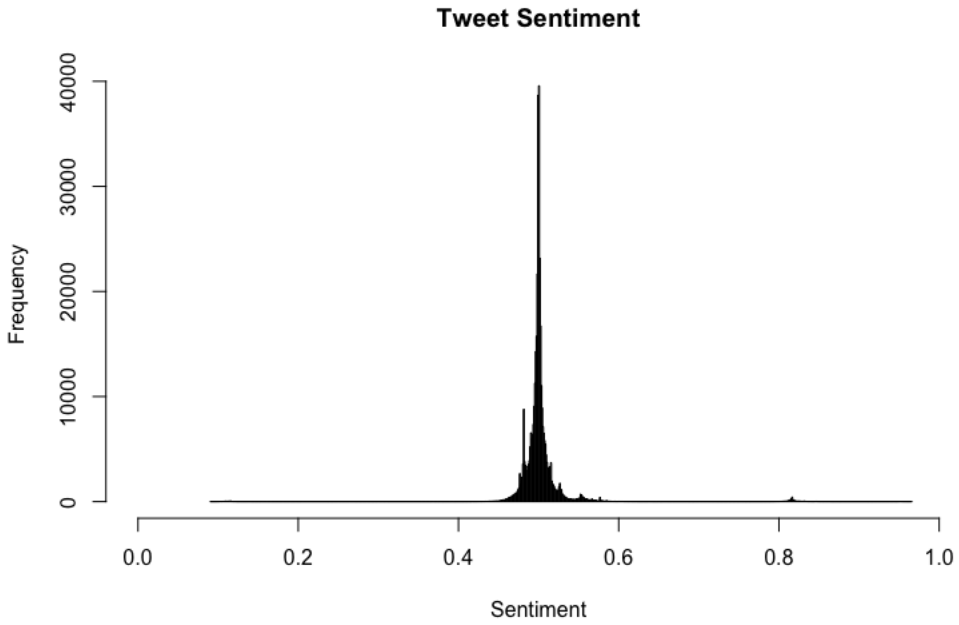

Figure 2: Histogram of Overall Sentiment

4.4.1 Sentiment Across Search Terms

One important question about consumer sentiment towards genome editing in domestic livestock is how consumers feel towards the words that are typically associated with genome editing. This is an important question because consumers may have different feelings towards the individual topics, terms or words that are associated with a general topic. For example, perhaps individuals have positive feelings towards GMO's in their food because GMO foods are typically less expensive per unit in the market, but they are specifically against gene editing because they believe that the editing process is inhumane towards animals. This example also shows that individuals can be uninformed about certain topics, and don't know exactly how the terms "GMO" and "gene editing" are related, and this could be something that educators or marketers work towards fixing by disseminating clearer information or advertisement campaigns.

Table 1 presents the summary statistics for the sentiment of each search term used to procure the data, and I find that each term has a different level of sentiment associated with it. The terms "organic" and "animal welfare" have the highest sentiment with 0.505 and 0.504 respectively, and the terms with the lowest sentiment are "dehorning" and "genetically modified" with a sentiment of 0.497 and 0.498 respectively. The difference in the search term's sentiment shows that individuals have differing views about the terms that are associated with genome editing in domestic livestock, and that while individuals may perceive "genome editing" as something positive, they also might perceive "genetically modified" and/or "GMO" as something negative. Another interesting thing about the predicted sentiment values for each search term is that the distance between the 25th and 75th percentiles seem to all be similar. These similar distances imply that the middle 50% of the distributions for each search term are similar in shape, which will help when I formally compare the distributions in Section 5.

Search Term	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
animal welfare	3823	0.504	0.036	0.107	0.496	0.505	0.880
biotechnology	2881	0.503	0.022	0.436	0.496	0.505	0.819
crispr	4009	0.502	0.022	0.418	0.496	0.503	0.844
dairy	138385	0.499	0.039	0.091	0.490	0.502	0.958
dehorning	73	0.497	0.010	0.472	0.496	0.500	0.542
gene editing	1166	0.501	0.016	0.419	0.497	0.503	0.820
genetically modified	2700	0.498	0.025	0.111	0.494	0.501	0.839
genome editing	193	0.501	0.010	0.464	0.499	0.504	0.553
GMO	21864	0.500	0.028	0.103	0.495	0.502	0.855
organic	209358	0.505	0.034	0.092	0.497	0.506	0.966

Table 1: Summary Statistics of Genome Editing Sentiment on Twitter by Search Term

4.4.2 Sentiment Across States

Another factor that might contribute to the sentiment about genome editing and the terms related to it is location in the Unites States, and more specifically, the state in which a tweet came from. I present two maps of the United States that show the average sentiment (Figure 3 (a)) and level of sentiment uncertainty (Figure 3 (b)) across the United States with regards to genome editing in domestic livestock and the terms that surround that topic it.[3] Results show that Oregon, South Dakota and Montana have the highest average sentiment over the past 10 years with a sentiment level of 0.503, 0.501, and 0.501, respectively. Wisconsin, Alaska and Texas have the lowest sentiment towards the topic of genome editing in domestic livestock with a sentiment level of 0.498, 0.498, and 0.498, respectively. Another interesting result is that many of the coastal parts United States have an average sentiment greater than the middle part of the United States. With a specific focus on the sentiment level in the western part of the US, it is apparent that the more populated areas in the US are more fond of the topics that surround genome editing. Lastly, the uncertainty towards genome editing and the related terms has mixed results with Wyoming having the lowest uncertainty and Rhode Island having the highest uncertainty. Something else to note is that the states that have large metropolitan areas seem to have less uncertainty in their sentiment.

5 Comparisons Between the Categorical Predictors of Sentiment

The prediction results from Section 4.4 are informative, but it is still not wise to make inference on these predicted sentiment values. In this section, I propose a method for comparing and taking inference on the factors that influence sentiment. Employing this method is important because it will enables me to say that there is a statistically significant difference between the distributions of two factors, and what percentage of the time this statistically significant difference arises. I first present the generalized version of the proposed methodology in Section 5.1, and then apply it to

[3]Note that I am keeping all search terms for this analysis as I think it is pertinent to analyze genome editing in domestic livestock and the conversation that surrounds it. This is important because it is often hypothesized that individuals have a difficult time disentangling the true difference between these terms, so analyzing them together should give a reasonable prediction on how individuals feel about the entire topic of genome editing.

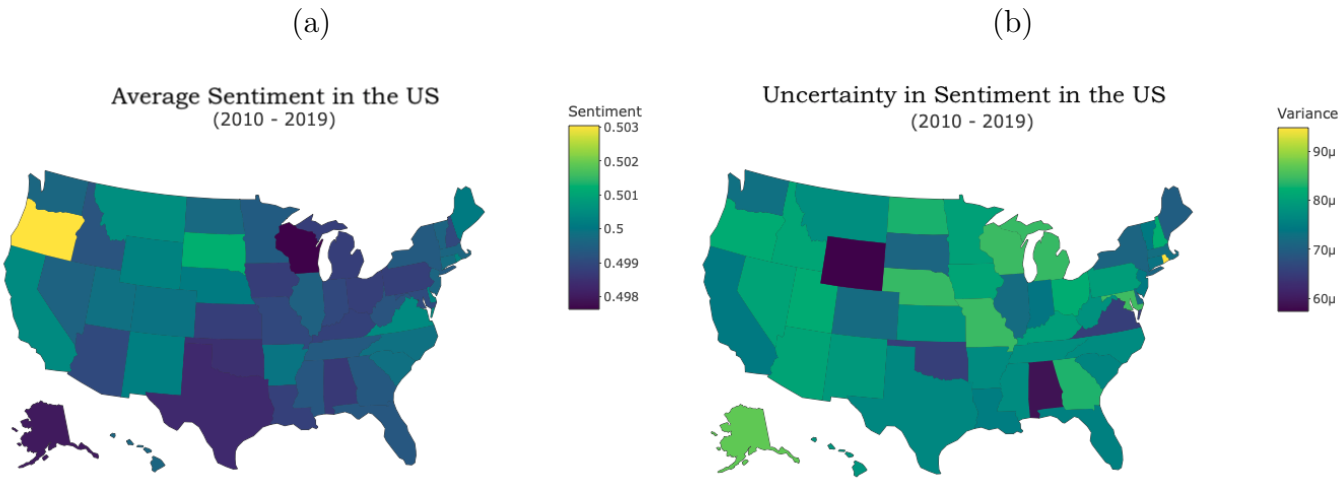

Figure 3: Sentiment in the United States for all Tweets

both the "search term" and "states" factor sets, from Sections 4.4.1 and 4.4.2, in Sections 5.2.1 and 5.3, respectively.

5.1 Statistical Model 2 - A One-Way ANOVA Model

To compare the means within each factor set, or each factor level pair, I employ a One-Way ANOVA model. The model takes the form of

$$y_{ij} = \mu_j + \epsilon_{ij} \tag{1}$$

for $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, g\}$, where $y_{ij} \in [0, 1]$ is the sentiment $\hat{y}_i$ mapped onto the sentiment index, $\mu_j \in \mathbb{R}$ is the real valued population mean for the j^{th} factor level, $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathbb{N}(0, \sigma^2)$ is a Gaussian error term, n_j is the number of observations in the j^{th} factor level and $n = \sum_{j=1}^{g} n_j$, g is the number of factor levels, and this formulation implies that $y_{ij} \overset{\text{ind}}{\sim} \mathbb{N}(\mu_j, \sigma^2)$.[4] The hypothesis test generated from the One-Way ANOVA model is

$$\begin{aligned} \mathrm{H}_0 &: \ \mu_1 = \cdots = \mu_j = \cdots = \mu_g \\ \mathrm{H}_A &: \ \text{At least one } \mu_j \text{ is different from the others} \end{aligned}$$

which is often rejected since each factor group has many different levels, and we would expect that *at least one* factor mean differs from the rest. In order to take inference from this hypothesis test, three assumptions about the residuals need to be verified. These assumptions are that the residuals have equal variance, come from a normal distribution and that there is no presence of auto-correlation. To test these assumptions, I first use the entire dataset with the One-Way ANOVA, but find that these assumptions are not satisfied with the exception of the no auto-correlation assumption.[5] This is not surprising because it is usually good practice to apply a One-Way ANOVA on data that has an equal sample size for each factor since it decreases the power and increases the Type I error

[4] Note that this model is equivalent to that of the dummy variable encoded general linear model $y_{ij} = \beta_0 + \sum_{j=1}^{g-1} \beta_j x_{ij} + \epsilon_{ij}$ where $\beta_0 = \mu_g$ and $\beta_j = \mu_j - \mu_g$ for $j \in \{1, \ldots, g-1\}$.

[5] I ran this initial analysis on both the "search terms" and "states" factor subsets.

rates in equivalence tests (Rusticus and Lovato (2014)). Additionally, it is obvious the predicted sentiment index $\hat{y}_i$ is not being generated from a normal distribution, and therefore, the Shapiro-Wilk's Test for Normality in the residuals is not appropriate for this data. This implies that standard parametric inference methods, such as Tukey's Method for Multiple Comparisons, cannot be used. These realizations lead me to implement a post-hoc non-parametric bootstrapping technique to find samples that first satisfy the equal variance assumption, and then test for significant differences between the means using a generalization of the Kruskal and Wallis (1952) test.

5.2 Bootstrap and Non-Parametric Diagnostic Analysis to Gain Inference

I bootstrap the entire dataset by randomly selecting 50 observations from each factor 10,000 times. This produces 10,000 samples, with 500 observation when sampling by search term and 2,500 by state, that I then use to run 10,000 independent Levene Equal Variance tests with an alpha significance level of $\alpha = 0.05$. For each Levene's test, we want to *fail to reject the null*, which happens approximately 70.69% of the time for the "search terms" factor subset and 82.67% for the "states" factor subset. This leaves 7,069 "search term" samples and 8,267 "state" samples that qualify for the non-parametric difference in means test. Figure 4 shows the distribution of the Levene's test p-value for both factor groups, and it is apparent that there is a high percentage of observations that are over the p-value = 0.05, the red vertical line. Using this subset of data, I employ a Wilcoxon–Mann–Whitney test (Wilcoxon (1945), Mann and Whitney (1947)) and control for the false discovery rate, to make the test more powerful, using the Benjamini and Hochberg (1995) adjustment.[6] This test compares all pair-wise combinations in each factor and tests if there is a statistically significant stochastic difference between them, and this test is used as a generalization of the difference in means test when running a post-hoc analysis on a One-Way ANOVA. Specifically, the hypothesis test is stated as:

$$\begin{aligned} \mathrm{H}_0 &: f(x) = g(x) \\ \mathrm{H}_A &: f(x) \neq g(x) \ \ \forall \, x \end{aligned}$$

where $f(x)$ and $g(x)$ are the probability distributions for each factor j. This test has also been used to implicitly compare the medians in each group j, but for this analysis, I keep it in general form. I present the results for both the "search term" and "state" factor groups in the next section, and refer to all of the means generated from these "qualifying samples" as "qualifying means." I use means to motivate the differences between the distributions since the mean has appealing bootstrapping properties.

[6] This test has many other names such as Mann-Whitney U Test, but I chose to call it the Wilcoxon–Mann–Whitney test to respect all contributions to the literature.

(a) (b)

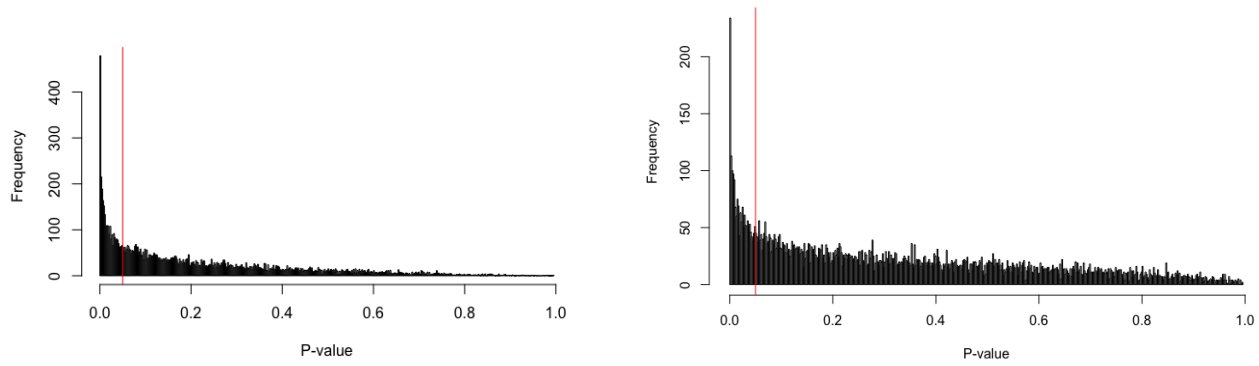

Figure 4: Histogram of Levene's Test P-values for Search Terms (a) and States (b)

5.2.1 The Difference in Sentiment Between Search Terms

I present the proportion of samples that reject the null that there is no difference between for the top 12 highest proportions in Table 2. I also present the mean of all sample's means in Table 2, and plot the qualifying bootstrapped means, the means that were rejected in the test, in Figure 5 for visual inspection. It is clear from Table 2 and Figure 5 that some of stochastic distributional comparisons are statistically different from each other and others are statistically different from each other more often than others. Overall, the term "dehorning" seems to always be perceived as negative and statistically different from many of the other terms. The terms "dehorning" and "genome editing" are the terms that are statistically different from each other the most. This is clear from the two different means presented in Table 2, and it is apparent in Figure 5 (a) that both distributions are significantly different. This intuition follows suit for the terms "genetically modified" and "genome editing" where there is a clear negative sentiment towards the term "genetically modified" and a positive sentiment towards "genome editing." This is an interesting result because some hypothesize that consumers do not perceive these terms differently, but 42.3% of the time, when all term's residuals are seen to be equal, these terms distributions are seen to be statistically different. To motivate this result even further, this implies that approximately 29.19% of the 10,000 samples statistically perceive "genetically modified" to be negative with a mean sentiment of 0.497 and significantly different than the positively perceived "genome editing" term with a mean of 0.502.

Search Term	Proportion Rejected	Mean 1	Mean 2
dehorning & genome editing	0.578	0.496	0.502
genetically modified & genome editing	0.413	0.497	0.502
dehorning & organic	0.359	0.496	0.506
dairy & genome editing	0.342	0.497	0.502
biotechnology & dehorning	0.322	0.504	0.496
dehorning & gene editing	0.288	0.496	0.502
animal welfare & dehorning	0.272	0.505	0.496
crispr & dehorning	0.241	0.503	0.496
genetically modified & organic	0.232	0.496	0.507
dairy & organic	0.229	0.496	0.506
biotechnology & dairy	0.205	0.504	0.496
biotechnology & genetically modified	0.203	0.505	0.496

Table 2: Top 12 Proportion of Wilcoxon–Mann–Whitney Test Rejections for Search Terms

I present all of the results for the 45 term comparisons in Section 7, the Appendix, and leave the rest of the interpretations to the reader. With that said, these results are very much for a particular hypothesis question applied to a subset of the population. For example, the highest proportion of rejected samples is the comparison between the terms "dehorning" and "genome editing" at 42.3%, which means that 40.86% of the total samples will see these terms as significantly different from each other, but this also implies that I can not compare 59.14% of the population's preference between these means. Since this comparison is the highest statistically significant proportion, I will proceed as assuming that these words serve as a subset for the topic of genome editing, and that the entire dataset can be seen as representing "the domestic consumers that have opinions on genome editing in livestock," and that their estimated sentiment is their opinion towards "genome editing in domestic livestock."

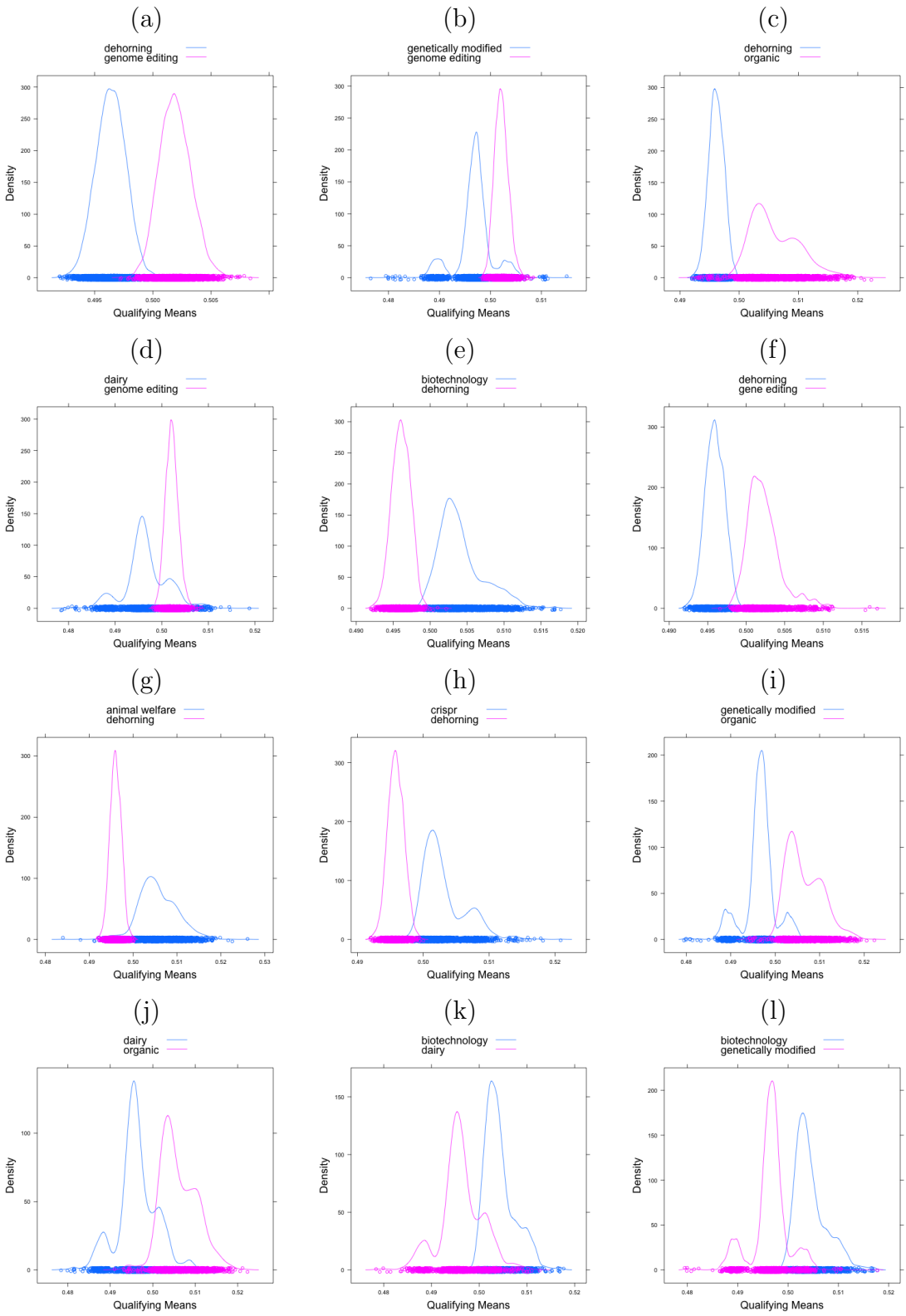

Figure 5: Comparing the Difference in Qualifying Mean Distributions

5.3 The Difference in Sentiment Between State

Table 3 presents the 12 factor comparisons that have the highest proportion of rejecting the null of the Wilcoxon–Mann–Whitney test out of the 8,267 samples that were drawn by state. Oregon had the highest sentiment overall from Section 4.4.2, and it also has a sentiment distribution that is greater than and stochastically dominates many of the other distributions. This is an interesting result because, for example, I can now say that 13.56% of the time ($0.8267 \times 0.164 \approx 0.1356$), Oregon and Oklahoma have significantly different sentiment distributions, and that the means of these sentiment distributions are 0.507 and 0.496, respectively. To further motivate these results, I plot the "qualifying means" in Figure 6 and find that even though many of them are bimodal, they look similar and there is a clear separation between the two distributions. Out of the top 12 highest proportion rejected, it is not surprising that Oregon, the state with the highest sentiment towards genome editing and the terms that surround this topic, has a sentiment distribution that is higher and significantly different than many of the other states because the states Oregon is being compared to have low sentiment and uncertainty levels. On the contrary, Oklahoma seems to have the some of the lowest sentiment and uncertainty towards genome editing and the terms that surround it, and this becomes apparent when being compared to the states like South Dakota and Maine, which have high sentiment and relatively lower uncertainty. With that said, the sentiment distributions between South Dakota and Oklahoma look similar, so taking inference is advised, but the distributions between Maine and Oklahoma are less alike, so caution is advised when taking inference. All other comparison interpretations are left to the reader, and I do not provide all comparisons in the Appendix as there would be a table of 1225 comparisons.

State	Proportion Rejected	Mean 1	Mean 2
OK & OR	0.164	0.497	0.507
OH & OR	0.122	0.496	0.507
OR & WV	0.117	0.507	0.497
AK & OR	0.116	0.497	0.507
AL & OR	0.102	0.498	0.507
KS & OR	0.096	0.498	0.508
OR & TX	0.088	0.508	0.497
LA & OR	0.085	0.497	0.508
OK & SD	0.080	0.496	0.504
ME & OK	0.074	0.514	0.496
MS & OR	0.070	0.500	0.508
KY & OR	0.069	0.497	0.508

Table 3: Top 12 Proportion of Wilcoxon–Mann–Whitney Test Rejections for States

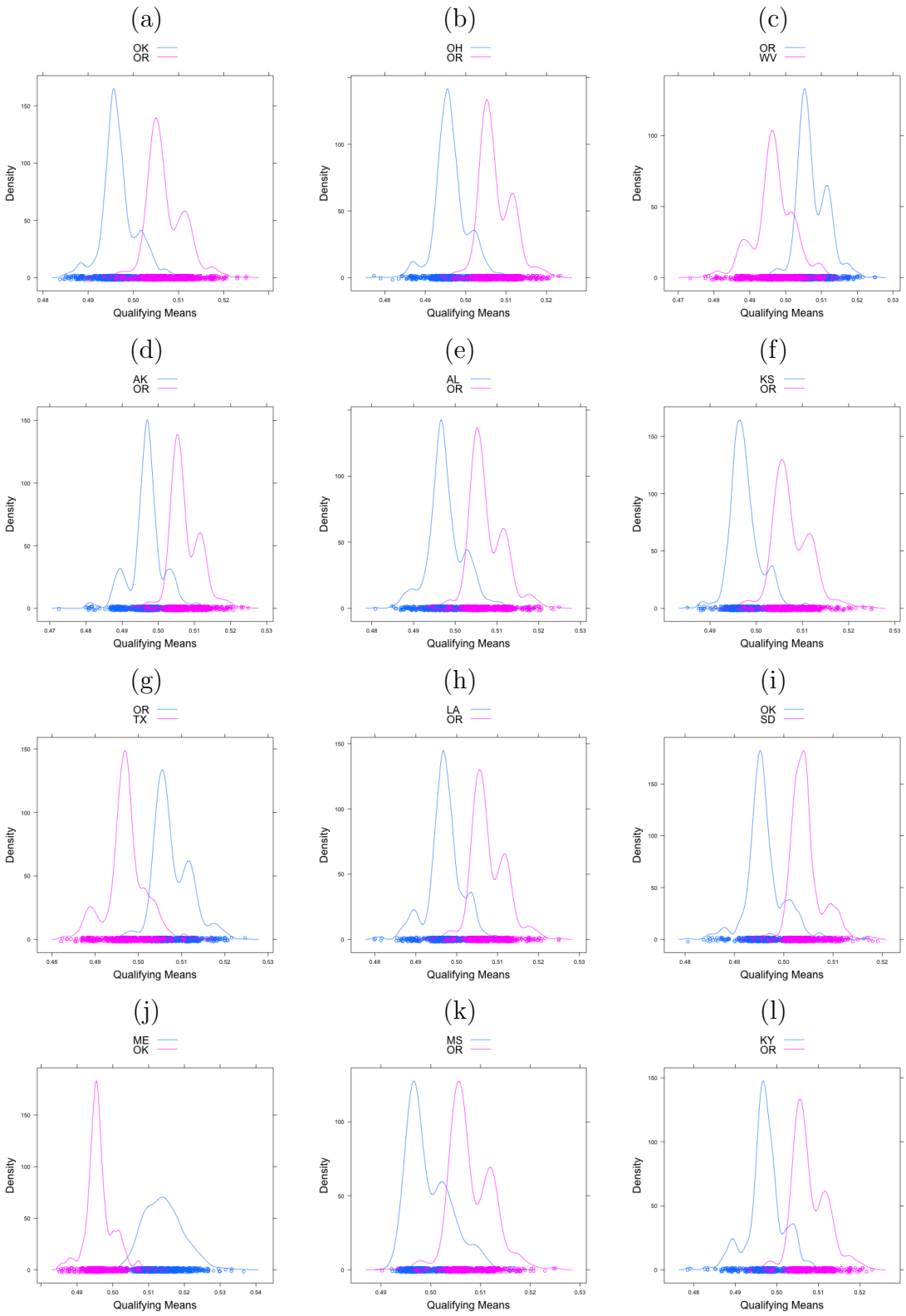

Figure 6: Comparing the Difference in Qualifying Mean Distributions

6 Conclusion

I use Twitter text data from the United States and machine learning to develop a sentiment index about genome editing in domestic livestock and the terms related to genome editing. I find that the sentiment towards the term "genome editing" in the United States is 0.501, and that the other terms that one would think are closely related to "genome editing," such as "biotechnology," "crispr," and "gene editing," all have similar sentiment level of 0.503, 0.502, and 0.501, respectively. The other related terms have mixed results in terms of sentiment, with the terms "dehorning" and "genetically modified" having the lowest sentiment at 0.497 and 0.498, respectively. This is an interesting result because it shows that Twitter users do have different sentiments towards different terms related to genome editing, and that even though genome editing, gene editing and biotechnology would be considered things that fall under the topic of genetic modification, they are perceived as something positive wile genetically modified is perceived as something negative. Policy makers should take this into account when regulating the labels on food and/or informing the public about what the difference is between these practices.

I also compare the overall sentiment for all the terms related to genome editing in domestic livestock by state, and find that Oregon, South Dakota and Montana have the highest overall sentiment towards these terms, and that Wisconsin, Alaska and Texas have the lowest sentiment towards these terms. I present the sentiment levels across all states in Figure 3 (a) and find that many of the states in the center of the contiguous United States, from Texas up to to Michigan, seem to have a lower sentiment levels than the coastal states. This is an important result because it is one of the first sentiment indexes related to genome editing in domestic livestock created for the United States. Policy makers and firms can use these sentiment levels to focus on the states that are more pro or anti genome editing in domestic livestock depending on the intent of their policy.

Lastly, I develop a methodology in order to gain inference from the sentiment predictions by bootstrapping the Twitter data based on the search terms to collect the data and the states in which the data came from. This analysis yields intuitive results in which I can take statistical inference on a proportion of the population. Many results are generated and are left for interpretation to the reader, but one main result answers the question of "do Twitter users perceive the terms related to "genome editing" as significantly different from each other?" The answer is yes, but only a proportion of the population sees them as statistically different. Two terms the population feels are significantly different are "dehorning" and "genome editing," and this happens 40.86% of the time with differing mean sentiments of 0.496 and 0.502, respectively. This implies that 40.86% the population has a statistically different sentiment level of 0.006 towards the words "dehorning" and "genome editing," with term "dehorning" being perceived as negative and "genome editing" being perceived as positive. This is an important result because it shows that a large proportion of the United States see these words as significantly different. This is just one of the results presented in this analysis, and readers are referred back to Section 5 for an exhaustive explanation of results. I also apply this analysis across states, and find interesting results. Policy makers can use these results to regulate labeling in consumer products by being more detailed about the practices that are happening in livestock production. For example, if policy makers allow genome editing in dairy production to reduce dehorning practices, there is a large part of the population that would accept this practice leading to an increase demand lowering production costs for farmers that currently engage in the expensive dehorning process.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Hopkins, D. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Ortega, D. L., Lin, W., and Ward, P. S. (2022). Consumer acceptance of gene-edited food products in china. *Food Quality and Preference*, 95:104374.

Rusticus, S. and Lovato, C. (2014). Impact of sample size and variability on the power and type i error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation*, 19(11).

Tabei, Y., Shimura, S., Kwon, Y., Itaka, S., and Fukino, N. (2020). Analyzing twitter conversation on genome-edited foods and their labeling in japan. *Frontiers in plant science*, 11:535764–535764.

van Rijsbergen, C., Robertson, S., and Porter, M. (1980). New models in probabilistic information retrieval.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wirz, C. D., Howell, E. L., Brossard, D., Xenos, M. A., and Scheufele, D. A. (2021). The state of gmos on social media: An analysis of state-level variables and discourse on twitter in the united states. *Politics and the Life Sciences*, 40(1):40–55.

7 Appendix

Compared Terms	Proportion Rejected	Mean 1	Mean 2
dehorning_and_genome editing	0.578	0.496	0.502
genetically modified_and_genome editing	0.413	0.497	0.502
dehorning_and_organic	0.359	0.496	0.506
dairy_and_genome editing	0.342	0.497	0.502
biotechnology_and_dehorning	0.322	0.504	0.496
dehorning_and_gene editing	0.288	0.496	0.502
animal welfare_and_dehorning	0.272	0.505	0.496
crispr_and_dehorning	0.241	0.503	0.496
genetically modified_and_organic	0.232	0.496	0.507
dairy_and_organic	0.229	0.496	0.506
biotechnology_and_dairy	0.205	0.504	0.496
biotechnology_and_genetically modified	0.203	0.505	0.496
gene editing_and_genetically modified	0.171	0.503	0.496
dairy_and_gene editing	0.171	0.496	0.502
animal welfare_and_dairy	0.168	0.506	0.496
genome editing_and_GMO	0.165	0.503	0.498
animal welfare_and_genetically modified	0.163	0.506	0.497
crispr_and_dairy	0.137	0.503	0.496
crispr_and_genetically modified	0.122	0.504	0.496
GMO_and_organic	0.092	0.498	0.507
biotechnology_and_GMO	0.072	0.505	0.498
dehorning_and_GMO	0.071	0.495	0.503
animal welfare_and_GMO	0.055	0.507	0.498
gene editing_and_GMO	0.050	0.503	0.497
crispr_and_genome editing	0.044	0.500	0.503
dairy_and_GMO	0.041	0.495	0.503
crispr_and_GMO	0.034	0.504	0.498
gene editing_and_genome editing	0.029	0.499	0.503
genetically modified_and_GMO	0.027	0.495	0.503
crispr_and_organic	0.025	0.500	0.508

Compared Terms	Proportion Rejected	Mean 1	Mean 2
dehorning_and_genetically modified	0.022	0.495	0.500
animal welfare_and_genome editing	0.019	0.500	0.503
gene editing_and_organic	0.019	0.500	0.508
biotechnology_and_crispr	0.017	0.506	0.500
animal welfare_and_crispr	0.014	0.508	0.499
dairy_and_genetically modified	0.013	0.495	0.499
animal welfare_and_organic	0.013	0.500	0.508
biotechnology_and_organic	0.013	0.502	0.506
crispr_and_gene editing	0.012	0.500	0.503
biotechnology_and_gene editing	0.011	0.505	0.500
biotechnology_and_genome editing	0.011	0.500	0.503
genome editing_and_organic	0.011	0.502	0.503
animal welfare_and_gene editing	0.011	0.506	0.500
dairy_and_dehorning	0.010	0.500	0.496
animal welfare_and_biotechnology	0.010	0.502	0.504

Compared States	Proportion Rejected	Mean 1	Mean 2
OK_and_OR	0.164	0.497	0.507
OH_and_OR	0.122	0.496	0.507
OR_and_WV	0.117	0.507	0.497
AK_and_OR	0.116	0.497	0.507
AL_and_OR	0.102	0.498	0.507
KS_and_OR	0.096	0.498	0.508
OR_and_TX	0.088	0.508	0.497
LA_and_OR	0.085	0.497	0.508
OK_and_SD	0.080	0.496	0.504
ME_and_OK	0.074	0.514	0.496
MS_and_OR	0.070	0.500	0.508
KY_and_OR	0.069	0.497	0.508
IN_and_OR	0.066	0.498	0.508
MI_and_OR	0.065	0.498	0.508
OR_and_PA	0.063	0.508	0.498
NE_and_OR	0.062	0.498	0.508
ME_and_OH	0.059	0.514	0.495
ND_and_OR	0.054	0.497	0.508
IL_and_OR	0.053	0.498	0.508
ME_and_WV	0.051	0.515	0.496

Compared States	Proportion Rejected	Mean 1	Mean 2
OH_and_SD	0.051	0.495	0.505
SD_and_WV	0.049	0.505	0.495
MO_and_OR	0.047	0.499	0.508
AK_and_SD	0.045	0.496	0.505
FL_and_OR	0.045	0.499	0.508
AR_and_OR	0.045	0.499	0.508
IA_and_OR	0.044	0.499	0.508
AK_and_ME	0.042	0.496	0.515
OR_and_WI	0.040	0.508	0.499
MD_and_OR	0.040	0.498	0.509
ID_and_OR	0.039	0.499	0.508
AL_and_ME	0.039	0.497	0.515
OR_and_WA	0.038	0.509	0.500
AL_and_SD	0.037	0.497	0.505
KS_and_ME	0.036	0.497	0.515
OK_and_VT	0.036	0.495	0.506
KS_and_SD	0.035	0.497	0.505
OR_and_VA	0.035	0.508	0.498
GA_and_OR	0.035	0.499	0.508
ME_and_TX	0.033	0.516	0.496
NM_and_OR	0.032	0.499	0.508
OR_and_WY	0.032	0.508	0.500
DE_and_OK	0.031	0.505	0.495
SD_and_TX	0.030	0.505	0.496
OR_and_SC	0.028	0.508	0.498
NH_and_OR	0.028	0.499	0.509
LA_and_ME	0.027	0.496	0.515
CA_and_OK	0.027	0.507	0.495
OR_and_RI	0.027	0.508	0.499
MN_and_OR	0.026	0.500	0.509