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

Washington State University
School of Economic Sciences

April 19, 2022

Introduction

Motivation:

Uncovering consumer sentiment towards genome editing in domestic livestock is difficult, and usually involves costly experiments or surveys.

Natural Language Processing (NLP) and Social Media Data provide alternative ways to gain insights about user, and perhaps consumer, sentiment.

These insights can help policy makers move forward on more representative policies.

- **Example Hypothesis:**
  If genome edited livestock products were to be introduced into the domestic market, how would Twitter users perceive these products and the key terms on the labels of the product. How would each state perceive these products?

Methods:

I use machine learning, with Twitter text data from the United States, to predict user sentiment about the term "genome editing" and the terms that surround the topic of genome editing in domestic livestock.

I investigate how user sentiment differs by "search term" and "state" factors.

Literature Review

There are three studies that are closely related to my research:

- **Ortega et al. (2022)** - Study how consumers accept genome edited pork products in China through a geographically dispersed consumer acceptance choice experiment and survey. Results show 38% of the respondents were in support of consuming gene-edited pork to prevent African Swine Flu (ASF), and that 30% of the respondents were in support of transgenetic pork to prevent ASF.
- **Tabei et al. (2020)** - Use Twitter data from 14,066 users to analyze their sentiment towards genome-edited foods and the labeling policy of Japan's Consumer Affairs Agency. Concluded that 54.5% to 62.8% of the tweets were negative about the Consumer Affairs Agency's labeling policy towards genome-edited foods.
- **Wirz et al. (2021)** - Use 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 non-parametric content analysis software, known as *Crimson Hexagon ForSight*, to predict the sentiment of tweets (Hopkins and King (2010). 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.

Contribution(s)

To my knowledge, this is the first study to predict the sentiment for terms that surround the specific topic of "genome editing in domestic livestock."

It is the first study to predict sentiment using Twitter's full archive from January 2010 to January 2021.

It is the first study to use tweets in which location was derived from the user's self-proclaimed location status.

It is the first to use text data to predict and gain inference from sentiment at different factor levels.

Outline

A Sentiment Analysis Using Twitter Data

- Identifying Tweets Related to a Particular Topic
- Social Media Data
- Statistical Model 1 - Sentiment Prediction with Logistic Regression
- Results
  - Sentiment Across Search Terms
  - Sentiment Across States

Comparisons Between Sentiment Factors

- Statistical Model 2 - A One-Way ANOVA Model
- Bootstrap and Non-Parametric Analysis to Gain Inference
  - The Difference in Sentiment Between Search Terms
  - The Difference in Sentiment Between State

Identifying Tweets Related to a Particular Topic

I first identify a set of keywords that relate to the topic of "genome editing in domestic livestock" using Social Mention.

The words that are most closely related to the topic of "genome editing in domestic livestock" are:

- **animal welfare**
- **biotechnology**
- **crispr**
- **dairy**
- **dehorning**
- **gene editing**
- **genetically modified**
- **genome editing**
- **GMO**
- **organic**

Social Media Data

I collect every tweet sent in the US from January 2010 to January 2021 that contains one of the key words aforementioned in Slide 6.

I code a simple text classifier that recognizes and imputes the state abbreviation for each tweet. This process reduces the dataset from 2 million to 384,452 observations.

(a) (b)

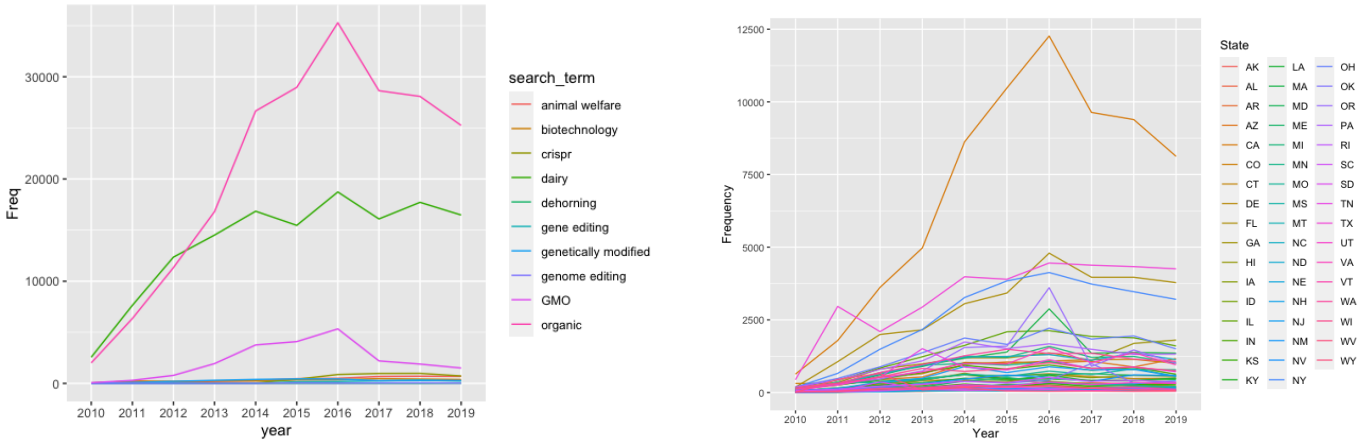

Figure 1: Frequency of Tweets in the United States about Genome Editing

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. The 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

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

- m is the number of training examples
- $y^{(i)}$ is the actual label of the i^{th} training example
- $h(z(\theta)^{(i)})$, or $\hat{y}_i$, is the model's prediction for the i^{th} training example

I use Python's Natural Language Tool Kit `NLTK` data to train the logistic regression (Bird et al., (2009)). The data has 10,000 tweets: 5,000 positive and 5,000 negative. I use the Porter stemming algorithm to stem words (van Rijsbergen et al. (1980)). With 8,000 training observations, the trained model has a 98% classification rate.

Preliminary Results

Twitter users have an overall sentiment towards "genome editing in domestic livestock" of 0.502, which is positive, and a standard deviation is 0.035.

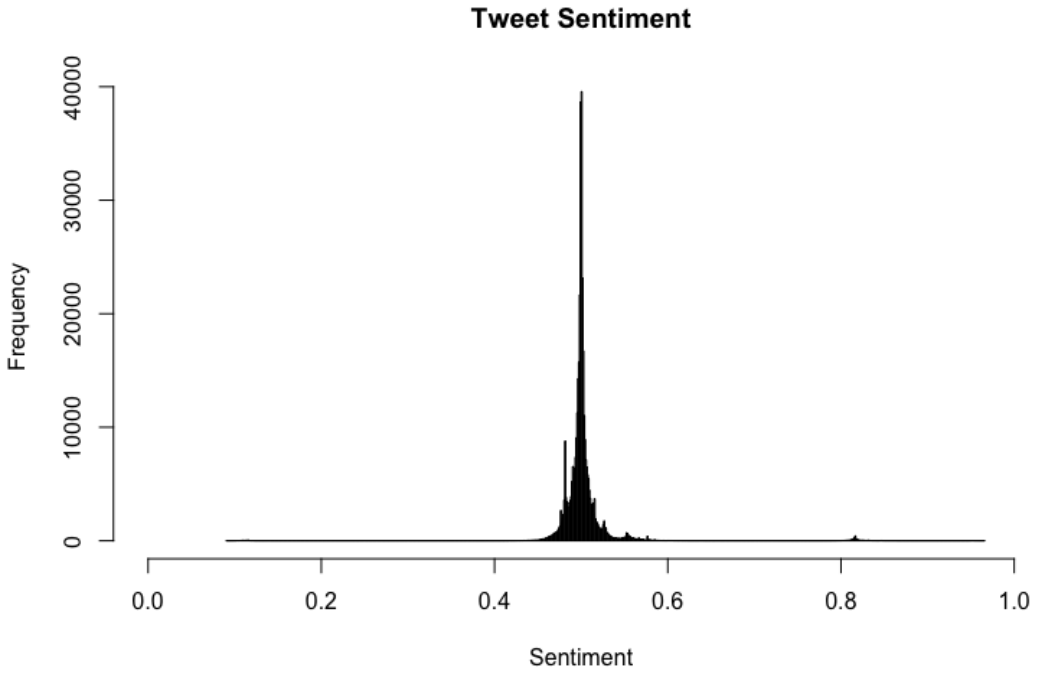

Figure 2: Histogram of Overall Sentiment

Sentiment Across Search Terms

Table 1 presents the summary statistics for the sentiment of each search term used to procure the data.

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 Sentiment on Twitter by Search Term

Summary of Results for Sentiment Across Search Terms

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 terms closely related to “genome editing,” such as “biotechnology,” “crispr,” and “gene editing,” all have similar positive sentiment levels of 0.503, 0.502, and 0.501, respectively.

Sentiment Across States

Figure 3 (a) shows the average sentiment and Figure 3 (b) shows the level of sentiment uncertainty in the United States. This is for all words related to "genome editing in domestic livestock."

(a) (b)

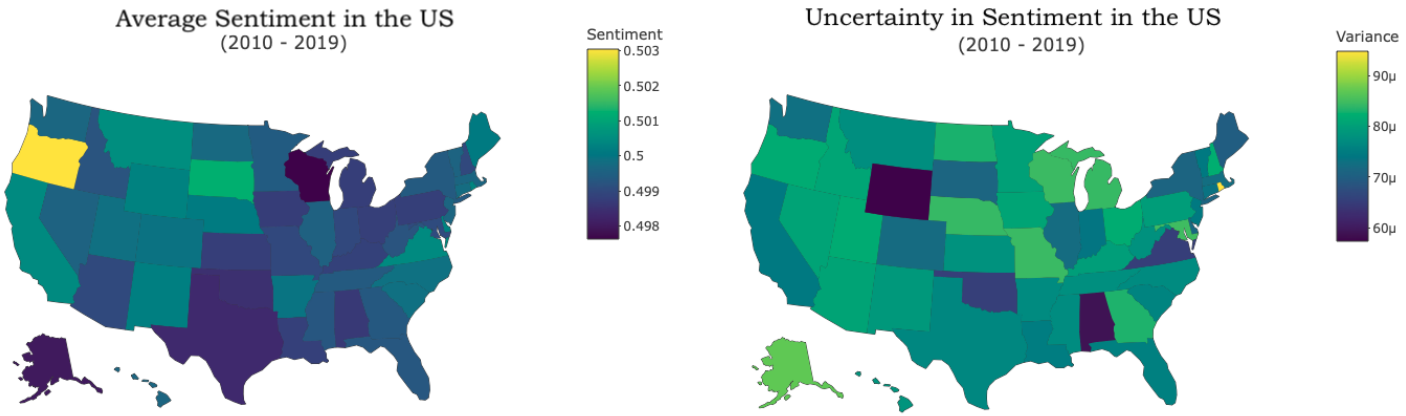

Figure 3: Sentiment in the United States for all Tweets

Summary of Results for Sentiment Across States

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.

On the contrary, 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.

Statistical Model 2 - A One-Way ANOVA Model

Compare the means within each factor group using a One-Way ANOVA model:

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

for $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, g\}$, where

- $y_{ij} \in [0, 1]$ is the sentiment 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

The hypothesis test generated from the One-Way ANOVA model is:

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

Statistical Model 2 - A QUICK NOTE

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 $j \in \{1, \ldots, g-1\}$.

We can also formulate this into a differences in the means model.

Statistical Model 2 - Inference Requirements

The hypothesis test generated from the One-Way ANOVA model is:

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

Hypothesis is often rejected since each factor group has many different levels. Use a post-hoc analysis to compare mean pairs (i.e the $\frac{g(g-1)}{2}$ factor level pairs).

Assumptions are that the residuals have equal variance, come from a normal distribution and that there is no presence of auto-correlation.

Assumptions are **not satisfied** with full dataset. Cannot do post-hoc analysis.

These diagnostic results lead me to implement a post-hoc bootstrapping technique to find samples that first satisfy the equal variance assumption, and then use a non-parametric test for a significant difference between the stochastic distributions of all factor pairs.

Bootstrap and Non-Parametric Analysis for Inference

Bootstrap the entire dataset by randomly selecting 50 observations from each factor level 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 $\alpha = 0.05$.

Levene's Hypothesis Test:

$$\mathrm{H}_0 : \ \sigma_1^2 = \cdots = \sigma_j^2 = \cdots = \sigma_g^2$$

$$\mathrm{H}_A : \ \text{At least one } \sigma^2 \text{ is different from the others}$$

Search Terms (a) States (b)

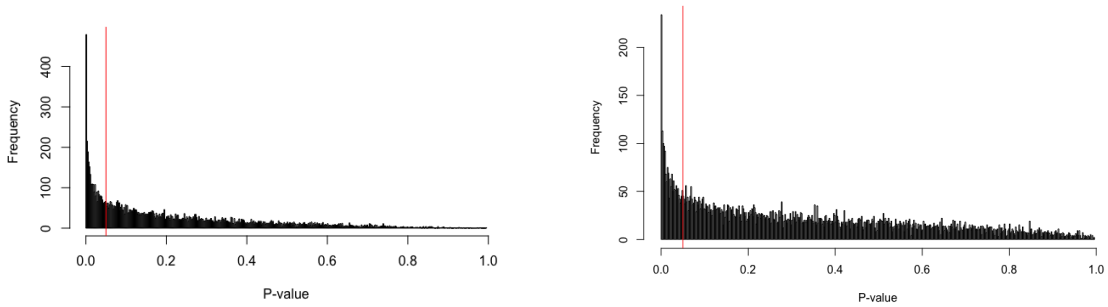

Figure 4: Histogram of Levene's Test P-values

Using this subset of data, I employ a Wilcoxon–Mann–Whitney test (Wilcoxon (1945), Mann and Whitney (1974)) and control for the false discovery rate, to make the test more powerful, using the Benjamini and Hochberg (1995) adjustment.

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.

The Wilcoxon–Mann–Whitney 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.

The Difference in Sentiment Between Search Terms

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 Test Rejections for Search Terms

The Difference in Sentiment Between Search Terms

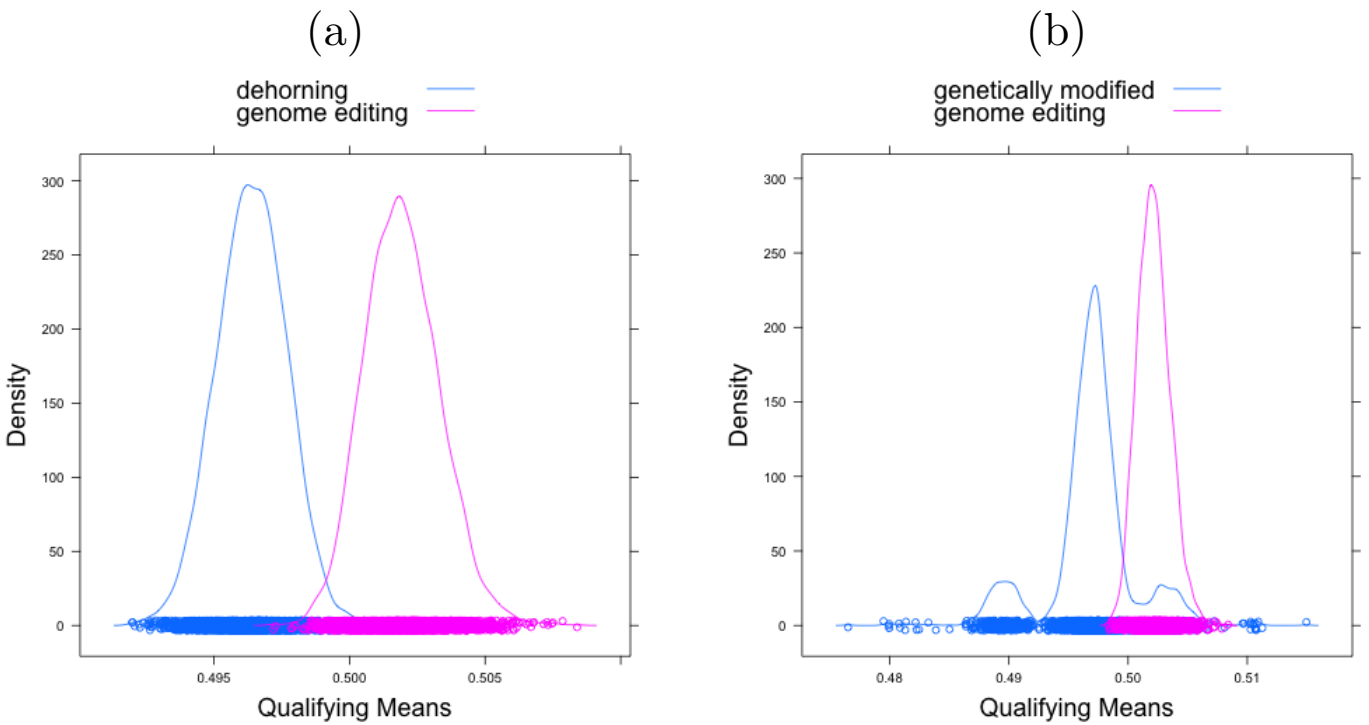

Figure 5: Histograms of Qualifying Sentiment Means Between Search Terms

The Difference in Sentiment Between State

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 Test Rejections for States

The Difference in Sentiment Between State

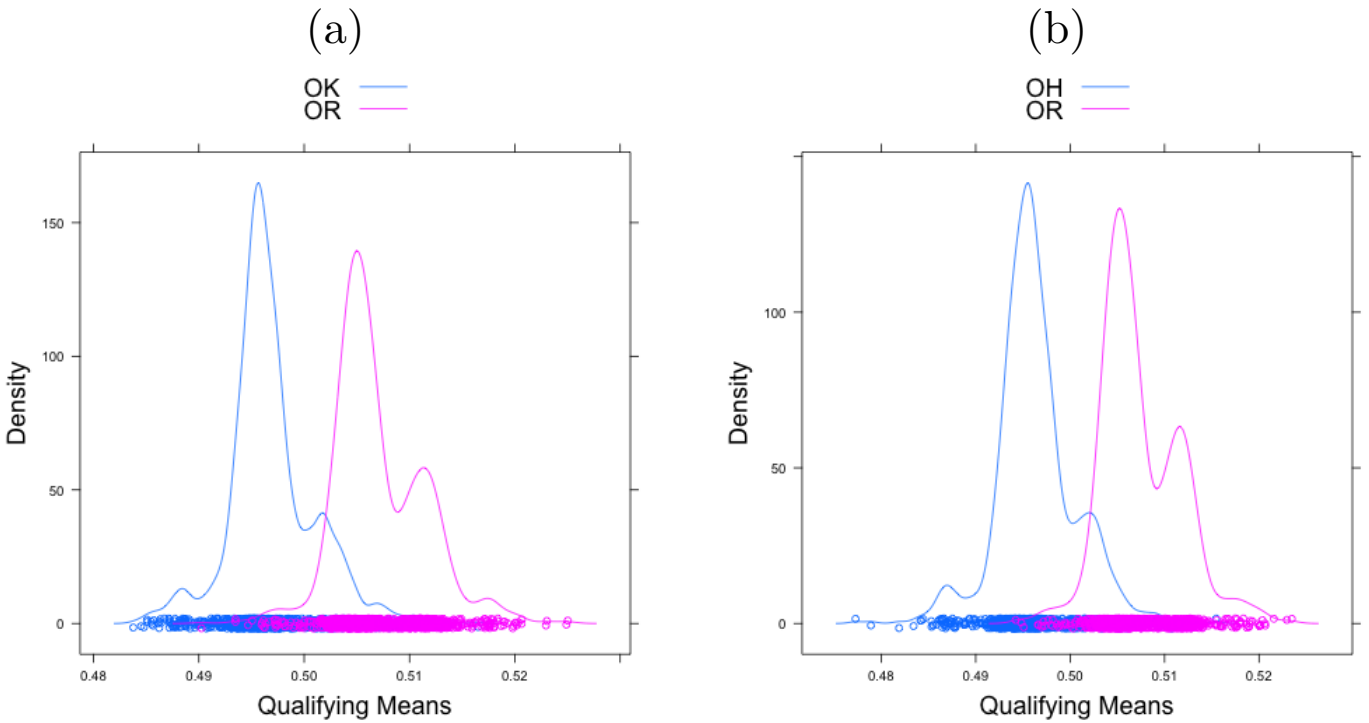

Figure 6: Histograms of Qualifying Sentiment Means Between States

Conclusions(s)

Twitter users have an overall sentiment towards "genome editing in domestic livesstock" of 0.502, which is positive.

The terms closely related to "genome editing," such as "biotechnology," "crispr," and "gene editing," all have similar positive sentiment levels of 0.503, 0.502, and 0.501, respectively.

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.

If policy makers want to introduce genome edited livestock products into the food market, they should be conscious on how they should label these products, along with non-genome edited products, because consumers on Twitter may perceive them differently.