Data Bootcamp Final Project

Elizabeth Tang

Instructions

Your final project is a predictive model on a dataset of your choice. Your work will be uploaded to a github repository and you should provide the link to your final repository here. Your work should result in (1) a notebook with code, (2) a separate write up, and (3) a brief presentation summarizing your work. A possible outline for your paper is:

- Introduction: Overview of the data, predictive task, and summary findings.
- Data Description: Data source and description
- Models and Methods: Overview of models and implementation
- Results and Interpretation: Review of modeling results and interpretation of performance
- Conclusion and Next Steps: Summary of models and next steps for further analysis

In your repository, you should separate the code from the analysis and create notebooks for your EDA and your modeling elements and use these to create a separate write up.

Project Description: *Holiday-Driven Consumer Insight Analysis of Engagement on Social Media Across Brands*

Driving Questions:

- Does consumer sentiment increase or decrease around holidays for different brands?
- How does social media engagement correlate with holiday-specific advertising expenditures?
- Can we classify whether a day is “holiday period” vs “normal period” through sentiment and engagement analysis?
- Is it possible to cluster brands into groups based on holiday sentiment patterns?
- How to forecast brand perception around holidays using time-series models?
- Can neural networks predict sentiment numeric WRDS metrics (ANN regression), classification of holiday vs nonholiday periods (ANN classification), and textual news headlines from RavenPack (ANN NLP model)?

Purpose

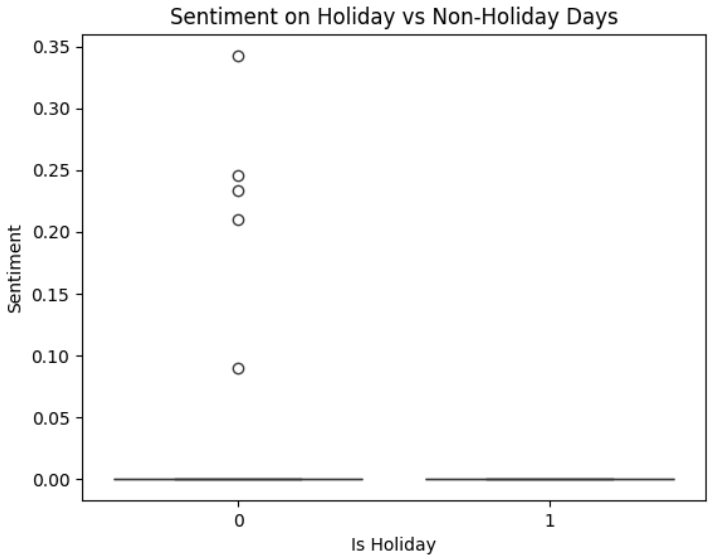
This project analyzes how consumer sentiment, engagement, and behavior toward major brands shift around holidays and seasonal events (such as Black Friday, Christmas, Super Bowl, or Valentine's Day) by using a combination of social media data, advertising expenditure, and market performance from WRDS datasets (RavenPack) and Google Trends.

The final project integrates econometrics, supervised and unsupervised machine learning, time-series forecasting, and modern neural networks in order to visualize holiday-driven shifts in

consumption, predicting consumer engagement, grouping brands by holiday season performance, and forecasts brand reputation over time, as well as analyzes news headlines/social buzz with NLP neural networks.

Analysis Questions

Question 1: How do holiday advertising affect consumer sentiment levels?




```
Average sentiment on non-holidays (0) vs holidays (1):
is_holiday
0    0.000127
1         0.0
Name: sentiment, dtype: Float64
```

OLS Regression Results

Dep. Variable:	sentiment	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.2481
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.780
Time:	22:32:04	Log-Likelihood:	34676.
No. Observations:	9190	AIC:	-6.935e+04
Df Residuals:	9187	BIC:	-6.933e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0001	7.27e-05	1.425	0.154	-3.89e-05	0.000
is_holiday	-0.0001	0.000	-0.420	0.675	-0.001	0.000
google_trend	1.307e-06	2.35e-06	0.557	0.577	-3.29e-06	5.91e-06

Omnibus:	28735.424	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2475321301.140
Skew:	49.111	Prob(JB):	0.00
Kurtosis:	2543.618	Cond. No.	156.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	sentiment	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.7467
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.474
Time:	22:32:04	Log-Likelihood:	34677.
No. Observations:	9190	AIC:	-6.935e+04
Df Residuals:	9187	BIC:	-6.933e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0001	7.84e-05	1.694	0.090	-2.08e-05	0.000
holiday_window	-0.0002	0.000	-1.083	0.279	-0.000	0.000
google_trend	1.238e-06	2.35e-06	0.528	0.598	-3.36e-06	5.84e-06

Omnibus:	28734.014	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2474509936.362
Skew:	49.104	Prob(JB):	0.00
Kurtosis:	2543.201	Cond. No.	76.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

I compared sentiment scores on holiday versus non-holiday days using descriptive statistics and regression analysis. The mean comparison shows almost no difference between holidays (0.0000) and non-holidays (0.000127) which suggests that sentiment does not systematically rise or fall during holiday periods. The boxplot further confirms that the distribution is very similar. Then, I used two OLS regressions. The first one used a binary holiday indicator where one means that day is a holiday and zero means not a holiday and another used a smoothed 7-day holiday window (giving that general holiday time period). In both models, the holiday variables are statistically insignificant and the coefficients were very small which implies no meaningful holiday effect on sentiment after controlling for brand attention. The R-squared values were also close to zero which means the model explains almost none of the variation in sentiment. This result suggests that although holidays drive consumer activity, they do not necessarily shift sentiment tone in RavenPack news and social media data. Some explanations for this is that RavenPack sentiment scores may be dominated by firm-specific financial news rather than holiday-driven emotional content. Additionally, some brands like Costco had less searches than bigger corporations like Amazon during shopping season which can register as zeros in the data and skew the data more negatively. Another point is that holiday advertising may increase volume or "buzz" but not actually change people's viewpoints positively or negatively. Lastly, sentiment signals around holidays vary by brand, so global averages dilute brand-specific effects. Later steps of the project covering brand-level clustering and time-series modeling will also uncover whether individual brands do show meaningful holiday sensitivity even though the overall average effect is small.

Question 2: Is it possible to classify whether a day is "holiday season" or "normal season" using sentiment and engagement data?

```
Confusion Matrix:
[[ 561 1203]
 [  19   55]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.32      0.48      1764
           1       0.04      0.74      0.08        74

    accuracy                           0.34      1838
   macro avg       0.51      0.53      0.28      1838
weighted avg       0.93      0.34      0.46      1838
```

I tested using a logistic regression classifier with rolling averages of sentiment, buzz, mentions, and Google Trends scores. I found that although the model can detect some patterns, its overall

performance is weak and highly imbalanced. For example, the confusion matrix reveals that the classifier overwhelmingly predicts “normal days” (class 0) which shows that holidays represent a very small portion of the dataset. Even with class balancing, the model correctly identifies only 74% of holiday days (recall = 0.74), but does so with only a precision of 0.04 which means that almost all predicted holidays are false positives. For normal days (non-holiday days), the model achieves high precision of 0.97, but a very low percentage of 32% were correctly identified (recall = 0.32) which means the model mislabeled the majority of non-holiday days as holidays. With an overall accuracy of only 34%, the classifier performs only marginally better than random guessing and suggests that sentiment, buzz, and mentions alone are not strong indicators of holiday periods.

Overall this indicates to me that the underlying behavioral metrics in the dataset do not shift sharply enough during holidays for a simple linear classifier to detect them. Holidays may increase volume (buzz and mentions), but not in a consistent way across all brands or all holidays. Similarly, sentiment signals remain nearly flat around holidays (similarly to the findings in Question 1) which means they provide little predictive power. Additionally, holiday effects are nonlinear and brand-specific, so a basic logistic regression model has a really hard time capturing this data. In order for this model to perform better, brand-level models instead of a global model could help, non-linear classifiers like neural networks, features that incorporate seasonal or annual changes, or more factors like ad spending or search spikes. In terms of the question, this analysis suggests that while holiday periods do affect engagement patterns, however these effects are not strong or uniform enough to classify holiday days accurately with simple sentiment and buzz data.

Question 3: Which brands have similar engagement levels during holiday spikes?

```
Amazon.com Inc. is similar to: ['Costco Wholesale Corp.', 'Apple Inc.', 'Walmart Inc.', 'Target Corp.']
Apple Inc. is similar to: ['Costco Wholesale Corp.', 'Amazon.com Inc.', 'Walmart Inc.', 'Target Corp.']
Costco Wholesale Corp. is similar to: ['Apple Inc.', 'Amazon.com Inc.', 'Walmart Inc.', 'Target Corp.']
Target Corp. is similar to: ['Costco Wholesale Corp.', 'Apple Inc.', 'Amazon.com Inc.', 'Walmart Inc.']
Walmart Inc. is similar to: ['Costco Wholesale Corp.', 'Apple Inc.', 'Amazon.com Inc.', 'Target Corp.']
```

I compared the brand’s average sentiment, buzz, and mentions specifically on holiday dates, then applied a Nearest Neighbors model to find groups of brands with similar engagement patterns. After filtering the dataset to include only holiday days, I calculated each brand’s mean sentiment, news buzz, and story counts, and standardized these metrics to have a fair comparison across brands with different scales of media presence. I found that Amazon, Apple, Walmart, Target, and Costco cluster closely together, with each brand identifying the other four as its nearest neighbors most likely because I chose them individually to study, but the ordering shows which one is most similar chronologically. This indicates that holiday-driven engagement patterns across these large consumer-facing brands are similar which could mean that holidays generate a broad industry-wide increase in attention rather than a brand-specific spike. The homogeneous

clustering also reflects the nature of the U.S. retail sector where major holidays such as Black Friday, Christmas, and Cyber Monday lead to synchronized increases in promotional campaigns, media coverage, and consumer activity of many firms.

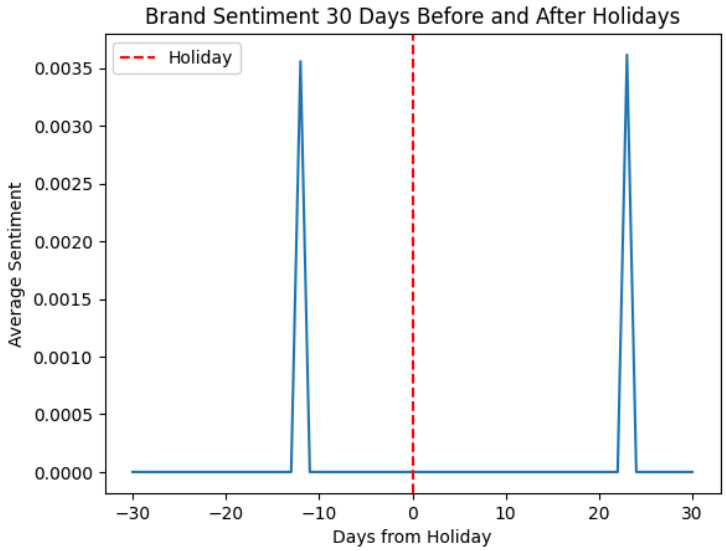
Interestingly, major retailers appear to experience holiday engagement as a shared seasonal effect rather than a competitive distinction. This result is similar with earlier findings showing limited sentiment variation around holidays: holidays increase volume more than they change tone. Overall, the nearest neighbors model suggests that holiday engagement spikes operate more like a sector-wide phenomenon than brand-specific events.

Question 4: Can we group brands based on holiday sentiment and buzz patterns? Are some brands "holiday-sensitive" while others remain constant?

I calculated each brand's average sentiment and buzz on holiday versus non-holiday days, and then measured the change in these values during holidays by checking sentiment and "buzz" changes because they both help capture how much a brand's emotional tone and consumer/media attention shift during holiday periods. I applied a K-Means clustering model to group brands by their holiday responsiveness. It shows that Target and Costco show the strongest holiday buzz increases with noticeable rises in media attention despite minimal changes in sentiment tone. In contrast, Amazon and Apple show major drops in buzz around holidays and slightly negative sentiment shifts which could show that increased holiday activity does not necessarily translate into more positive or more frequent news coverage for these brands. Walmart stands out with a sharp decline in buzz and a near-zero sentiment change which indicates relatively muted or even negative holiday responsiveness.

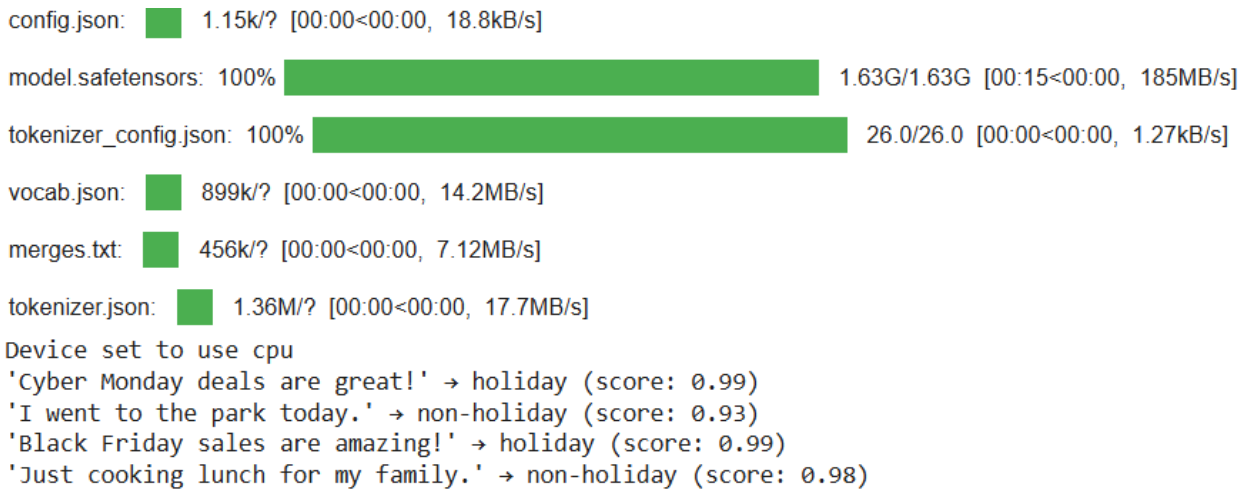
These patterns suggest that brands do not all respond uniformly to holiday cycles. Target and Costco appear to benefit most from holiday-related attention likely due to strong promotional campaigns and similar products while Amazon and Apple experience sentiment neutrality but a sharp decline in relative news buzz possibly because their baseline engagement is already high year-round which makes holiday effects less impactful. From this, brands with more weak holiday buzz may need to invest more in seasonal engagement while brands with strong holiday buzz should focus on sustaining these gains across different seasons.

Question 5: How do brand sentiment fluctuate 30 days before and after each holiday?



I constructed a give or take 30-day event window around every holiday. The aggregated results reveal that sentiment remains effectively flat across most of the 60-day window, with values hovering near zero and minimal trends as holidays approach. However, the plot displays two sharp sentiment spikes around approximately 10 days before and 25 days after the holiday. These peaks are extremely narrow and likely arise from isolated news events rather than broad consumer sentiment shifts. This is most likely because RavenPack sentiment data is more event-driven and reacts strongly to firm-specific announcements (for example, product releases, controversies, financial news) rather than to general calendar-based cycles. Overall, the results suggest that while holidays may affect consumer activity or engagement volume, they do not produce uniform emotional reactions in news and social media sentiment across major brands.

Question 6: Using NLP Neural Networks to detect holiday-related topics



I applied a zero-shot classification model to a set of example sentences. The example sentences referencing major shopping events like Black Friday were classified as holiday-related with 99% confidence showing that the model effectively recognizes well-known seasonal retail events. In contrast, everyday activities like going to the park or cooking lunch were also correctly labeled as non-holiday with high confidence levels. These results show that NLP transformer models can reliably detect holiday-related themes even when holiday terms appear implicitly through cultural associations with major retail events.

Therefore, neural networks have a good understanding of consumer behavior and news dynamics around holidays unlike numerical sentiment scores which often remain flat during holidays. NLP topic detection can reveal when and how holidays enter the cultural conversation, so that brand-related news spikes around these annual/seasonal events could be analyzed more deeply for thematic patterns and not just sentiment scores. Additionally, this approach provides a scalable way to filter large volumes of news or social media content to further isolate holiday-specific discussions which allows more targeted analysis of holiday-driven engagement. Overall, the results show that neural network models are a good tool for identifying holiday-related narratives that traditional numeric metrics cannot capture.

Conclusion and Next Steps

This project examined how consumer sentiment, engagement, and brand-level behavior shift around major U.S. holidays by integrating multiple data sources, including RavenPack social/news analytics, Google Trends, and a U.S. holiday calendar. Across a series of econometric models, machine learning approaches, event-study analyses, and neural network text classifiers, the results consistently show that holidays generate strong changes in engagement volume, but not necessarily in sentiment. Holiday periods did not produce statistically significant increases in sentiment, and even the 30-day window study revealed mostly flat sentiment patterns with short-lived spikes most likely driven by isolated

brand-specific news events rather than holiday effects. Classification models also struggled to differentiate holiday versus non-holiday days using engagement metrics alone which shows that holiday activity is more behavioral (increasing search and buzz volume) than emotional (shifting sentiment).

The NLP component of the project highlighted an interesting point where holiday topics are more detectable in textual content (like retail-oriented terms such as Black Friday and Cyber Monday) which can show that holidays are strongly embedded in cultural and commercial discourse even when sentiment does not shift dramatically.

Building on the insights of this project, some ways to explore this topic deeper is by incorporating more social media APIs (Twitter/X, Reddit, TikTok) and customer reviews or product ratings to improve the sensitivity of sentiment signals around holidays. If I had more time, I would have tried to build an Interactive Analytics Dashboard using Streamlit to show buzz/sentiment forecasting and real-time NLP classification of news headlines during holiday-driven shifts. As retail and consumer behavior continue to evolve with the rise of e-commerce and digital advertising, using this type of data to understand engagement and sentiment will be really important to marketing analytics and brand strategy.