Evaluating Twitter as an Agricultural Economics Research Tool

by

Candace Elaine Gatson Smart

B.S., University of Missouri, 2016

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Agriculture Economics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Dr. Glynn T. Tonsor

Copyright

Abstract

Over the past decade, social media has risen from an emerging novelty to the normative form of expression for many Americans. As these platforms have risen in popularity, researchers have recognized the potential for capturing information users are self-reporting about their beliefs and preferences. Simultaneously, social media corporations have become privy to the value of this information being freely shared by consumers and have safeguarded much of their historical data to monetize the data. Faced with both an enticing new source of data, but a steep price to obtain it, researchers must evaluate the potential gains that can be extracted from the often difficult to analyze data.

This study explored the acquisition of social media, namely Twitter, data and the potential uses in the field of agriculture economics. A contract was secured with Sysomos, a social media analytics firm, in July of 2017 to collect raw Twitter data over the proceeding thirteen months. Changes in frequency of tweets and sentiment scoring of tweets were used to attempt to explain election results from November 2017 proposed legislations pertaining to marijuana and minimum wage as well as to explain and predict changes in the stock prices of selected publicly traded firms in the food producing sector. Twitter frequency changes were then compared to changes in traditional print media articles in an effort to determine the exchangeability of the two media sources when used to track events pertaining to animal health.

Results of this study suggested that Twitter data possess little power to explain the studied election results, but creation of a strong model was difficult due to the limited number of months of data available. Changes in the frequency of tweets were not found to be a strong indicator of changes in the stock market on the average day, but were shown to explain potentially highly valued information to investors on days with large changes in price. Twitter

and traditional print media were shown to be unique sources of data when exploring the topic of animal health events.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to acknowledge my major professor, Dr. Glynn Tonsor. Dr. Tonsor has provided endless support, encouragement, and abounding patience throughout my time at Kansas State University. I cannot thank him enough for his dedication to seeing me succeed in this program.

I would also like to extend my thanks to the hard-working women of 348 Waters Hall for providing their assistance and expertise every step of the way. They make Waters Hall a happy place to work.

Dr. Joe Parcell, at the University of Missouri, took a chance on me as a budding economist and for that I am ever grateful. His guidance and support set the course for my eventual graduate career at KSU.

Finally, I would like to thank my husband, family, and friends for their earnest interest in my research and confidence in my abilities.

Chapter 1 - Introduction

Academic researchers, investors, and players within the food production industry have long attempted to predict the preferences of consumers utilizing consumption data, pricing information, demographics, and media coverage. Consumer sentiment information, a highly sought-after influencer of demand, has long proved difficult to collect and quantify.

In the mid-2000’s, social media sites became mainstream and users began sharing personal information, preferences, and opinions online, information important to many researchers. Ways to access and analyze this data are continually being developed. This study will focus on the Twitter platform and explore ways the data can inform variables in predictive and explanatory models relating to the food producing industry.

Twitter Demographics

In 2016, 23% of U.S. adults that use the internet were Twitter users at that time (Newberry, 2016). Of internet users, 30% of urban, 21% of suburban, and 15% of rural dwellers used Twitter. Around 32% of online 18 to 50 year olds, 13% of 50 to 64 year olds, and 13% of those 65 and older, were active Twitter users. 25% of male and 21% of female adults used the platform. At the time, mothers were 67% more likely than women without children to research products using Twitter and 45% more likely to purchase an item due to Twitter influence (Newberry, 2016).

Twitter Drawbacks as a Research Tool

Despite availability on the demographics of Twitter users as a whole, many tweets do not provide information about the income, gender, age, or location of the tweet author, disallowing researchers to accurately correct for the differences in Twitter demographics and population demographics.

Many of the studies that utilize Twitter data to explain or predict an action, rely on sentiment scoring, or the ability to summarize the opinion or message of the tweet into a single value. To assign sentiment values to the large quantity of tweets and eliminate the subjectivity of the researcher, natural language processing (NLP) techniques are used (Kanakaraj & Guddeti, 2015). Problematically, basic techniques are created for use with well-written, standard text, while many tweets use casual and slang language.

Twitter Strengths as a Research Tool

Twitter touts many advantages as a data source for researchers. While the restraints of traditional surveying often cap the sample size for a study, with 313 million monthly users, 65 million of whom were located in the U.S., publishing 500 million tweets per day in 2016, Twitter provides information about a vast number of individuals (Newberry, 2016). These users are voluntarily surveying on a number of topics with statements such as “Obama has my vote” and “Can’t wait to go see the new Avengers movie on Friday” providing information about their intended actions. Twitter allows users to create a type of “custom newspaper” by following the friends, family, politicians, celebrities, and news sources of their choosing. Simply viewing the list of accounts a user follows can provide information on the opinions, priorities, and preferences of the user. Not only is this information voluntarily supplied by users, it is being recorded in real time, providing researchers with predictive information about events such as elections, stock price changes, and product revenues, in the hours and minutes leading up to the event.

Objectives

This study will explore the value of Twitter data as a research tool in the following ways:

- Determine Twitter data's ability, coupled with demographic data, to predict voting outcomes in a specific region.
- Determine Twitter data and traditional print media's ability to explain own-stock price changes for food producing firms as well as predict out-of-sample firm price changes and the potential financial implication of making trading decisions based on Twitter data.
- Explore Twitter's explanative power compared to traditional print media's using the topic of animal health.

Motivation

This study is motivated by uncertainty of the value of Twitter data against the cost of collection and analyzation as social media data is ever increasingly commercialized. In order to determine the potential financial gains from utilizing twitter information in decisions within food production and investing, the predictive ability of the data must be determined.

Organization of Thesis

Seven chapters comprise this thesis. Chapter 2 will review a selection of existing literature discussing the use of consumer sentiment in forecasting models and social media data as a research tool. Chapter 3 outlines the collection of Twitter and traditional print media data. Chapter 4 discusses the ability of Twitter data to explain election results. Chapter 5 discusses the explanatory and predictive value of Twitter data in the stock market. Chapter 6 discusses Twitter's exchangeability with traditional print media as a data source. Chapter 7 provides conclusions and discussion of potential future research.

Chapter 2 - Literature Review

Consumer sentiment has long been used in models to explain and/or predict demand in the food industry, with this information being sourced from surveys, focus groups, panels, and traditional print media articles. Tonsor, Mintert, and Schroeder (2010), using a Rotterdam model, estimated U.S. meat demand from 1982 to 2007 using meat consumption, prices, and media information. The authors created an index to measure consumer interest in the Atkins diet, or other high-protein, low-carbohydrate programs, using newspaper articles. Articles were scored as either positive (promoting the diet) or negative (warning of adverse effects of the diet) to determine the net consumer sentiment over time. Unlike Tonsor, Mintert, and Schroeder's other indices used in the model, newspapers were utilized instead of medical journals to represent consumer opinion rather than medical opinion.

Since the introduction of Twitter in 2006, some researchers have begun utilizing Twitter information to capture the sentiment of consumers, in lieu or in addition to the previously mentioned sources. Batrinca and Treleaven (2014) review the methodology of social media research as well as the analyzation tools available. Published in 2014, the authors recognize that the social media data may become less readily available to those in academic research due to the potential financial value of the data, a prediction that has come to fruition in the years since publication. Three definitions provided by the authors relevant to this study are:

- **Scraping-** "collecting online data from social media and other Web sites in the form of unstructured text and also known as site scraping, web harvesting and web data extraction."

- **Opinion mining**- “opinion mining (sentiment mining, opinion/sentiment extraction) is the area of research that attempts to make automatic systems to determine human opinion from test written in natural language.”
- **Sentiment analysis**- “sentiment analysis refers to the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials.”

Batrinca and Treleaven identify many of the potential challenges academic researchers face when utilizing social media data including the expense to access complete data sets, the storage required for ever-expanding data, and the unstructured nature of the data that can lead to misinterpretation of tweets when processed through an algorithm.

Azar and Lo (2016) posed a key question: “Can we infer this information from more traditional sources, or is it truly *new* information?” The cost associated with Twitter data collection and analyzation requires there to be unique, valuable information found in tweets that cannot be found elsewhere. The authors collected tweets referencing the Federal Reserve to predict stock market changes following Federal Open Market Committee (FOMC) meetings. Tweets were sentiment scored using a Python package, resulting in being assigned a polarity score falling between -1 (purely negative) and +1 (purely positive). The authors felt the algorithm while largely accurate, misinterpreted the sentiment of some tweets. In addition to assigning a sentiment score to each tweet, the sentiment score was weighted using the tweeter’s number of followers, a measure of the tweet’s reach. Azar and Lo found that even tweets authored by those with little expertise about the stock market contain information that can be used to predict stock market changes and utilizing Twitter data to make portfolio decisions can result in a higher performing portfolio than one selected disregarding Twitter information. The authors recognized both the challenges of transforming tweets into usable variables and the predictive potential of tweets as a data source unique from existing sources.

Ranco, Aleksovski, Caldarelli, Grcar, and Mozetic (2015) found a low correlation between Twitter sentiment and stock prices over time, but found a significant correlation between the sentiment of tweets during tweet volume peaks and returns. The authors used the "event study" method to tie Twitter sentiment about the thirty firms that comprise the Dow Jones Industrial Average Index to specific events that cause Twitter volume on the stock to rise. These peaks in volume are considered "events" and are assigned a polarity (determined by the sentiment as either negative, neutral, or positive). The authors found large Twitter volume around the time of earnings announcements, but peaks also occurred at times not corresponding to an earnings announcement. The cumulative abnormal returns (CAR) for the stock increased after peaks of positive sentiment tweets and decrease after peaks of negative sentiment tweets. The authors recognized that their models did not forecast stock market changes, but rather explained historical changes using tweets, which while informative, is not as financially lucrative.

In addition to determining the opinion of Twitter users, there is potential for substantial value to firms to understand the intended actions and purchasing decisions of the consumers active on Twitter. The authors of *Predicting movie Box-office revenues by exploiting large-scale social media content* (Liu, Ding, Chen, Chen, & Guo, 2014) mined tweets containing the title of an upcoming movie to determine the user's "purchase intention" (plan to go see the movie in theaters). "Purchase intention" is then used to predict the box-office revenues for that film, with a strong correlation between the two being found. Tweets were additionally assigned a sentiment score of positive, negative, or neutral, though the authors found purchase intention to be a stronger predictor of revenues than the sentiment score. The authors highlight social media's value as a form of volunteer surveying. Rather than implement a costly and time sensitive survey to mine data on consumer's opinions and "purchase intent", users are self-reporting this information. Not only does Twitter provide information about consumers' intended actions, it also serves as an "electronic word of mouth", influencing the decisions and opinions of their followers.

Chapter 3 - Data Collection

Twitter Data Collection

To obtain Twitter data, a contract was formed with Sysomos, a social media analytics firm. Sysomos, based in Toronto, allows clients to build searches for words or phrases used in tweets, news articles, and other social media posts. For this project, only Twitter data was used. This project focuses on Twitter discussions and if Twitter content significantly differs from print media for the topics studied.

Access to the Sysomos platform was gained on July 10th, 2017 and ran through September 10th, 2017. Sysomos allows clients access to the past two years of content on most social media platforms; Twitter access is limited to the previous thirteen months due to the high volume of posts. Because searches were downloaded over the course of the access period, most searches were conducted to include results from August 1st of 2016 to July 1st of 2017.

Boolean search language was used to build the searches. Boolean logic allows the searcher to force or exclude certain relationships between words (Elmer E. Rasmuson Library, 2016). The language used in a Boolean search is always presented in all capital letters. AND and OR are the two main commands used. AND requires that two specified words or phrases must both appear in the content for a tweet or news article to appear in the search results. For example, a search of (“E. coli” AND “pork”) only yields results that feature both words, helping to focus the results to a specific livestock species. OR allows several similar words to be substituted for one another to create a more complete results set. For example, in the searches conducted, each time “E. coli” was a phrase searched for, the search was constructed as (“E. coli” OR “E coli” OR “ecoli”) to capture differences in stylization. Searches pertaining to poultry were conducted as (“poultry” OR “turkey” OR “chicken”) to include not only results mentioning “poultry” but also results mentioning the specific species that make up the category of “poultry”. A search designed to capture discussion about legal action was stylized as (“settlement” OR “lawsuit”) in attempt to capture two sets of results that were highly similar.

Searches were constructed to gain a pulse on the discussions surrounding food safety and food recalls, particularly those pertaining to *E. coli*. Topics in animal health were also studied. Results pertaining to state election results from November 2016 were downloaded to be used as a control for using Twitter activity to predict voting outcomes. Minimum wage and marijuana legalization legislation were selected for this use. A list of the searches conducted and downloaded can be found in the tables 3.1, 3.2, 3.3, and 3.4.

Sysomos offers numerous results files that can be downloaded as a CSV, as well as graphs and images. The “mentions” file includes the data on each tweet or news article that is used to create the other analyses that can be downloaded. A single mentions file can include up to 50,000 results for a particular search. Most of the searches conducted for this project fell below that limit. For those that exceeded the limit, a random sample of 5,000 tweets was downloaded. Each tweet included in the results of a search is listed on a unique row of the spreadsheet. Due to the time constraints of the contract, only the mentions and latest activity files for Twitter results and news results was downloaded for each search. The content of the mentions file will be outlined below. Most of the other files available for download from Sysomos, such as the “sentiment” file and the “demographics” file can be replicated using the data from the mentions file. The latest activity file contains a list of the number of tweets or news articles that were posted on each day of the selected time frame. This file can also be replicated using the mentions file, but was downloaded for each list due to its anticipated frequent use.

The following search will be used to identify the information included in each mentions file for Twitter data. The search was intended to capture conversation surrounding PEDV and its relationship with biosecurity.

Query	("PEDV" OR "Porcine epidemic diarrhea virus") AND ("biosecurity" OR "bio security")
Start Date	2016-08-11 00:00:00
End Date	2017-09-01 23:59:59
Filter	(~SOURCE~TWITTER)

“Query” lists the exact search entered into Sysomos’ platform. The timeframe of the search is also listed at the top of the file. The results are filtered to only include Tweets, rather than all social media activity. 127 results were found for the time period. The example below shows the first result of the search, which is recorded in a single row in the spreadsheet.

Link	Date(ET)	Time(ET)	LocalTime
http://twitter.com/MBSwineSeminar/statuses/901431272794451969	8/26/2017	9:08:57	8/26/2017 8:08

“Link” provides a URL that can be used to access the tweet on Twitter’s website. “Date(ET)” and “Time(ET)” provide the date and time at which the tweet was posted on Twitter by the tweeter. “Local Time” classifies the tweet by the time the tweet was posted in the time zone in which it was composed.

Author ID	Author Name	Author URL	Authority	Followers	Following
MBSwineSeminar	ManitobaSwineSeminar	http://twitter.com/mbswineseminar	4	248	572

"Author ID" is the handle of the tweeter and appears attached to the @ sign. For the above example, the Author ID would appear as "@MBSwineSeminar". This username is used when twitter users reply to one another's tweets or want to mention another user in a tweet. For this reason, each twitter username, sometimes also referred to as a handle, must be unique. The "Author Name" is the more formal name selected by a user that is used on their Twitter page and at the top of each tweet, but cannot be used in replies or mentions because it is not required to be unique. Some users have identical IDs and Names. "Author URL" is a link to the user's twitter profile. The URL is simply the author's username after the backslash following Twitter's address, hence the need for unique IDs. The "Authority" column is an influence ranking assigned by an algorithm created by Sysomos. The scale rates a user on a scale of 1-10, 10 being the most authoritative and is designed to be a more accurate measure of influence than merely a user's number of followers. The authority score takes into consideration the user's number of followers, number following, the frequency at which the user posts updates, and how many times the user's tweets are retweeted by others. Authority rankings are specific to a particular user and apply to all of their tweets and are therefore not a score of how influential the tweet in the results set was. News outlets, celebrities, and high-profile experts tend to have the highest authority scores. This score will allow for greater ability to identify the demographics of those who are tweeting about a particular topic, whether it be largely media sources or consumers who have little influence. The "Followers" and "Following" columns measure the number of other users the tweeter has following their profile and as well as the profiles the tweeter is following, respectively. The "Followers" tab is an indicator to help predict how many users viewed the tweet. The ratio of followers to following helps identify users who others find particularly influential.

Gender	Language	Country	Province/State	City	Location
	English	Canada	mb	winnipeg	Winnipeg, Manitoba, CANADA

“Gender” is an optional detail that users can choose to disclose on Twitter. For this example result set, 35 of the 127 tweets were posted by a user that has their gender disclosed on their profile. Some of the profiles without a gender are news outlet sources that post content from a number of authors. “Language” identifies the language in which the tweet was composed and posted. Tweets appear in their original language regardless of the language preference of the person accessing the content, although users have the option to have a translation displayed for a specific tweet. The location from which the tweet is posted is also data that is sourced from voluntary reporting by the user. Sysomos uses any reported data from the user as well as any information about location provided by the user in the tweet or their Twitter bio to determine the location. Not all tweets in the result set have an identified location. In the example set, 88 of 127 tweets included at least “Country” location. 78 provided “Providence/State”. 61 included “City” identification. The “Location” column provides the most specific location available. The example above identified the country, providence, and city, so the location is listed as “Winnipeg, Manitoba, Canada”. If only the providence and country had been identified, the location column would have read “Manitoba, Canada” and simply “Canada” if only the country had been identified. The country name is typically left off for the “Location” column for results located in the U.S. In some cases, a specific city is not identified but information is available to help identify a broader region. For example, in this set, one result listed “U.S.” for country and “IL” for state, but the location column was listed as “Northern Illinois”.

Sentiment	Snippet
POSITIVE	"@UMN_swine_group: Are biosecurity measures for personnel efficient in preventing PEDV transmission? https://t.co/Ig67Uh7ch1"

Contents
"@UMN_swine_group: Are biosecurity measures for personnel efficient in preventing PEDV transmission? https://t.co/Ig67Uh7ch1"

Summary
"@UMN_swine_group: Are biosecurity measures for personnel efficient in preventing PEDV transmission? . https://t.co/Ig67Uh7ch1"

Bio	Unique ID
Sharing ideas and information for efficient pork production for over 30 years...	9.01431E+17

“Sentiment” is scored as either POSITIVE, NEGATIVE, or NEUTRAL by a Sysomos algorithm. The language and tone of the tweet is analyzed to determine the sentiment about the topic held by the tweeter. As there is a date attached to each tweet, sentiment can be mapped over the time period to identify any shifts in sentiment.

The “Bio” column contains the entirety of the user’s twitter bio. The twitter bio is a short introduction (160 character limit) that a user creates for their profile to help characterize the nature of their tweets or their perspective. The bio can provide insight into whether the tweet is in the same vein as the user’s other tweet and if the user holds a particular expertise for the topic.

This example tweet comes from a user whose twitter account focuses on pork production. The rest of this results includes a few users who left their bio section blank, a number of individuals and news outlets focusing on pork related topics, and users identifying as farmers, fathers, journalists, mothers, and sports fans.

Due to the 160 character limit of tweets, there is little need to summarize or condense the results for downloading. Because of this, for Twitter results, the “Snippet”, “Contents”, and “Summary” columns provide the tweet in its entirety and are therefore duplicates of each other.

A “Unique ID” is assigned to each tweet in Sysomos’ platform.

Searches

A multitude of searches were created to capture information about the November 2016 election (used in Chapter 4 to explain voting outcomes), food safety, food producing firms (used in Chapter 5 to estimate stock price changes), and animal health (used to determine the exchangeability of Twitter and traditional news sources in Chapter 6).

Table 3.1 Voting Searches

Query	Date Range	Twitter Hits
Initiative 1433	2016-08-01 2017-07-01	550
Amendment 70	2016-08-01 2017-07-01	1,391
Prop 206	2016-08-01 2017-07-01	11,399
Amendment 70 AND vote no	2016-08-01 2017-07-01	73
Prop 206 AND vote no	2016-08-01 2017-07-01	102
Prop 206 AND vote yes	2016-08-01 2017-07-01	134
Minimum Wage AND Arizona	2016-08-01 2017-07-01	9,380
Minimum Wage AND Maine	2016-08-01	6,754

	2017-07-01	
Minimum Wage AND Colorado	2016-08-01 2017-07-01	5,077
Initiative 1433 AND vote yes	2016-08-01 2017-07-01	34
Amendment 70 AND vote yes	2016-08-01 2017-07-01	4,094
Referred Law 20 AND vote no	2016-08-01 2017-07-01	4
Minimum wage AND lower	2016-08-01 2017-07-01	50,773
Minimum Wage AND abolish	2016-08-01 2017-07-01	2,914
Minimum wage AND raise	2016-08-01 2017-07-01	220,569
Prop 205	2016-08-01 2017-07-01	17,082
Prop 205 AND vote yes	2016-08-01 2017-07-01	1,431
Prop 205 AND vote no	2016-08-01 2017-07-01	1,066
Prop 64 AND vote no	2016-08-01 2017-07-01	102
Prop 64 AND vote yes	2016-08-01 2017-07-01	134
Prop 64	2016-08-01 2017-07-01	11,399
Amendment 2	2016-08-01 2017-07-01	60,272
Issue 6	2016-08-01 2017-07-01	98,477
Issue 6 AND vote yes	2016-08-01 2017-07-01	153
Issue 6 AND vote no	2016-08-01 2017-07-01	70
Question 2 AND vote yes	2016-08-01 2017-07-01	1,928
Question 2 AND vote no	2016-08-01 2017-07-01	3,113

Question 2	2016-08-01 2017-07-01	229,663
Question 3	2016-08-01 2017-07-01	297,095
Question 3 AND vote yes	2016-08-01 2017-07-01	1,710
Question 3 AND vote no	2016-08-01 2017-07-01	386
Question 4	2016-08-01 2017-07-01	134,291
Question 4 AND vote yes	2016-08-01 2017-07-01	2,046
Question 4 AND vote no	2016-08-01 2017-07-01	583
I-182	2016-08-01 2017-07-01	1,415
I-182 AND vote yes	2016-08-01 2017-07-01	83
I-182 AND vote no	2016-08-01 2017-07-01	14
Measure 5	2016-08-01 2017-07-01	4,865
Measure 5 AND vote yes	2016-08-01 2017-07-01	59
Measure 5 AND vote no	2016-08-01 2017-07-01	2
Arkansas AND marijuana	2016-08-01 2017-07-01	37,072
Arizona AND marijuana	2016-08-01 2017-07-01	46,864
Florida AND marijuana	2016-08-01 2017-07-01	156,805
California AND marijuana	2016-08-01 2017-07-01	265,872
Massachusetts AND marijuana	2016-08-01 2017-07-01	96,673
Montana AND marijuana	2016-08-01 2017-07-01	14,579
North Dakota AND marijuana	2016-08-01	16,202

	2017-07-01	
Nevada AND marijuana	2016-08-01 2017-07-01	88,409
Amendment 70 AND vote no	2016-08-01 2017-07-01	73
Minimum Wage AND South Dakota	2016-08-01 2017-07-01	619
Minimum Wage AND Washington	2016-08-01 2017-07-01	7,479
Amendment 2 AND vote yes	2016-08-01 2017-07-01	4,094
Amendment 2 AND vote no	2016-08-01 2017-07-01	965
Referred Law 20	2016-08-01 2017-07-01	54

Table 3.2 Animal Health Searches

Query	Date Range	Twitter Hits
PEDV	2016-08-08 2017-09-01	1,569
HPAI	2016-08-08 2017-09-01	59,122
BSE	2016-08-06 2017-09-01	220,226
PRRS	2016-08-06 2017-09-01	16,190
Biosecurity	2016-08-08 2017-09-01	54,973
PEDV AND biosecurity	2016-08-11 2017-09-01	115
HPAI AND biosecurity	2016-08-11 2017-09-01	400

Table 3.3 Food Safety Searches

Query	Date Range	Twitter Hits
E. Coli AND fatal OR death	2016-08-01 2017-07-01	677

E. Coli AND settlement OR lawsuit	2016-08-01 2017-07-01	843
E. Coli AND sick	2016-08-01 2017-07-01	5,400
Food Safety AND E. Coli	2016-08-01 2017-07-01	913
Recall AND flour	2016-08-01 2017-07-01	11,929
Food Safety AND recall	2016-08-01 2017-07-01	40,858
Food Safety AND beef or veal	2016-08-01 2017-07-01	1,571
Food safety AND breach	2016-08-01 2017-07-01	6,754
Chipotle AND food safety	2016-08-01 2017-07-01	6,445
FSIS	2016-08-01 2017-07-01	5,085
Recall AND frozen beef	2016-08-01 2017-07-01	51
Recall AND pork or ham or bacon	2016-08-01 2017-07-01	4,993
Recall AND undeclared allergens	2016-08-01 2017-07-01	2,515
Recall AND E. Coli	2016-08-01 2017-07-01	11,194
Recall AND ground beef	2016-08-01 2017-07-01	911
Recall AND lamb or mutton	2016-08-01 2017-07-01	489
Recall AND poultry or chicken or turkey	2016-08-01 2017-07-01	20,339
Recall AND poultry	2016-08-01 2017-07-01	1,621
Recall AND pork	2016-08-01 2017-07-01	1,911
Recall AND turkey	2016-08-01 2017-07-01	4,496
Recall AND chicken	2016-08-01	14,740

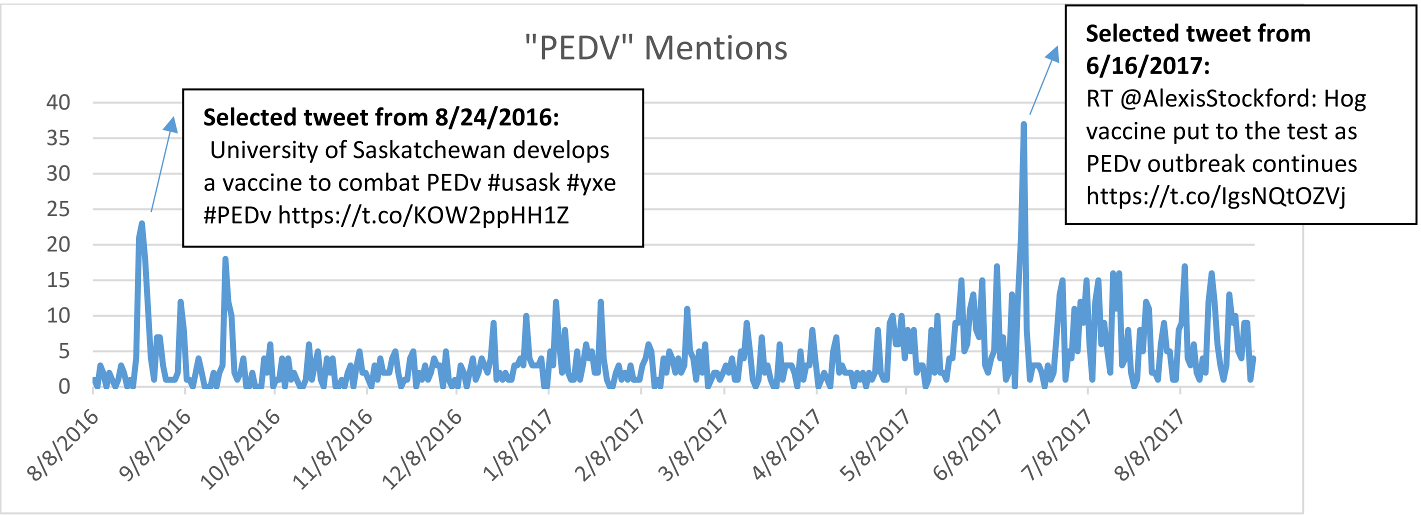
	2017-07-01	
E. Coli AND Chipotle	2016-08-01 2017-07-01	8,381
E. Coli	2016-08-01 2017-07-01	15,8321
Food Safety AND hazard	2016-08-01 2017-07-01	213
Food safety AND risk	2016-08-01 2017-07-01	3,169
Food safety AND policy	2016-08-01 2017-07-01	956
Food safety AND fear	2016-08-01 2017-07-01	99
Food safety AND scare	2016-08-01 2017-07-01	231
Food safety AND health risk	2016-08-01 2017-07-01	74
E. coli AND beef	2016-08-01 2017-07-01	8,146
Recall AND beef	2016-08-01 2017-07-01	11,631
Recall AND beef or veal	2016-08-01 2017-07-01	12,109

Table 3.4 Food Producing Firm Searches

Query	Date Range	Twitter Hits
ConAgra	2016-08-08 2017-09-01	23,184
Maple Leaf Foods	2016-08-08 2017-09-01	3,587
Cargill	2016-08-08 2017-09-01	139,385
Campbell Soup	2016-08-09 2017-09-01	144,250
ConAgra Recall	2016-08-09 2017-09-01	591

Maple Leaf Foods Recall	2016-08-09 2017-09-01	167
Cargill Recall	2016-08-09 2017-09-01	105
Campbell Soup Recall	2016-08-09 2017-09-01	531
Tyson Foods	2016-08-09 2017-09-01	66,686
Pilgrims Pride	2016-08-09 2017-09-01	4,652
Sanderson Farms	2016-08-11 2017-09-01	12,104
Seaboard Corporation	2016-08-11 2017-09-01	426
Cresud	2016-08-11 2017-09-01	3,323
BRF S.A.	2016-08-11 2017-09-01	1,059
Hormel Foods	2016-08-11 2017-09-01	11,114
Industrias Bachoco	2016-08-11 2017-09-01	422
Leucadia National	2016-08-11 2017-09-01	3,519
Aoxin Tianli	2016-08-11 2017-09-01	525

Figure 3.1 PEDV Daily Mentions Frequency and Selected Tweets



Selected tweet from 8/24/2016:
University of Saskatchewan develops a vaccine to combat PEDv #usask #yxe #PEDv https://t.co/KOW2ppHH1Z

Selected tweet from 6/16/2017:
RT @AlexisStockford: Hog vaccine put to the test as PEDv outbreak continues https://t.co/IgsNQtOZVj

Figure 3.1 provides the frequency of mentions over the search time period for a selected search, namely "PEDV or porcine epidemic diarrhea virus". A selected tweet from each of the two days of highest Twitter activity highlight the event that contributed to the frequency peak.

LexisNexis Data Collection

To compare Twitter coverage to traditional print media, news articles were accessed through NexisUni, the academic platform of LexisNexis. The "News" category includes newswires and press releases, industry trade press, newspapers, and web-based publications, among other news outlets. Due to the continual access to LexisNexis throughout this study, companion searches to the Twitter queries were completed as variables were selected for models. When used as a comparison to Twitter, searches were run using search terms and date ranges identical to the corresponding Twitter search. The results were downloaded for the time period, and the frequency of articles for each day was calculated. LexisNexis does not provide sentiment scoring for articles.

Chapter 4 - Twitter as an Explanatory Tool for Elections

2016 Election

On Tuesday November 8, 2016, eight U.S. states voted on measures to legalize marijuana and five states voted on measures to change minimum wages levels (Politico Staff, 2016). The marijuana initiatives all pertained to expanding the legalization of marijuana within the state, with some states voting to legalize medical use of marijuana and others recreational use. With the exception of Florida's Amendment 2, which required a 60% super-majority to pass, all other marijuana measures required a simple-majority to pass (Politico Staff, 2016).

Table 4.1 Marijuana Voting Measures

State	*Legislation*	*Outcome*
Arizona	Prop 205	Failed
Arkansas	Issue 6	Passed
California	Prop 64	Passed
Florida	Amendment 2	Passed
Massachusetts	Question 4	Passed
Montana	I-182	Passed
Nevada	Question 2	Passed
North Dakota	Measure 5	Passed

Table 4.2 Minimum Wage Voting Measures

State	*Legislation*	*Outcome*
Arizona	Prop 206	Passed
Colorado	Amendment 70	Passed
Maine	Question 4	Passed
Washington	Initiative 1433	Passed

Arizona, Colorado, and Maine voted to increase their minimum wage to $12 by 2020 and Washington's Initiative 1433 proposed an increase to $13.50 by 2020 (Politico Staff, 2016). All four measures required a simple-majority to pass, with all four states passing their minimum wage increases.

Data pertaining to South Dakota’s Referred Law 20, which proposed decreasing the minimum wage for those under 18 to $7.50 per hour, was not used due to a low volume of tweets on the subject and to being the only state proposing a reduction in minimum wage rather than a raise.

Twitter mentions downloaded from Sysomos were used as variables in models that estimated the percent of “yes” votes for the ballot issues. Searches were created and downloaded that tracked Twitter conversation pertaining to the name of each piece of proposed legislation, but due to some states having highly unique issue names, such as “I-182”, and others using more ambiguous terms such as “Amendment 2”, it was difficult to accurately identify conversations pertaining to the topic. Instead, the results were downloaded for each state name along with the topic of the issue, such as “Colorado AND minimum wage”.

The data was cleaned to only include results with a location tag within the state of the issue. Then, tweets without a city identified were eliminated. The cities were grouped into counties and the counties were combined to make between two and four regions per state, creating a total of 32 regions. The increase in observations from grouping by state to grouping by region allowed the degrees of freedom to increase. Additionally, this allowed the testing of whether media has a greater impact on voting results in some regions than in others. The regions were assigned as either rural or urban, which became a binary variable with Urban=1 meaning the region is comprised of urban counties and Urban=0 representing rural counties.

Tweets originating outside the state of the election were excluded. While these tweets may have influenced voters, the authors were likely not eligible voters and could not be tied to a voting outcome. In additional research, these out-of-state tweets could be used to create new variables to consider the influence of non-voters in the election. The exclusion of these observations greatly inhibited the sample size of the models. For example, 3,474 tweets were published on election day that contained mention of "Arkansas" and "marijuana", but only 463 of those tweets were identified as having originated within Arkansas. The sample size was reduced further as only 304 of the in-state tweets included a city location.

Models

Model 1

$$\_\_Yes = \beta_0 + \beta_1\_Freq\_ + \varepsilon$$

Model 2

$$\_\_Yes = \beta_0 + \beta_1 PercentPositive + \varepsilon$$

Model 3

$$\_\_Yes = \beta_0 + \beta_1 PercentNegative + \varepsilon$$

Model 4

$$\_\_Yes = \beta_0 + \beta_1 Marijuana + \varepsilon$$

Model 5

$$\_\_Yes = \beta_0 + \beta_1 Urban + \varepsilon$$

Model 6

$$\_\_Yes = \beta_0 + \beta_1\_Freq\_ + \beta_2 Marijuana + \varepsilon$$

Model 7

$$\_\_Yes = \beta_0 + \beta_1 PercentPositive + \beta_2 Marijuana + \varepsilon$$

Model 8

$$\_\_Yes = \beta_0 + \beta_1 PercentNegative + \beta_2 Marijuana + \varepsilon$$

Model 9

$$\_\_Yes = \beta_0 + \beta_1\_Freq\_ + \beta_2 Urban + \varepsilon$$

Model 10

$$(10)\ \_\_Yes = \beta_0 + \beta_1 PercentPositive + \beta_2 Urban + \varepsilon$$

Model 11

$$(11)\ \_\_Yes = \beta_0 + \beta_1 PercentNegative + \beta_2 Urban + \varepsilon$$

Model 12

$$(12)\ \_\_Yes = \beta_0 + \beta_1\_Freq\_ + \beta_2 PercentPositive + \beta_3 PercentNegative + \beta_4 Marijuana + \beta_5 Urban + \varepsilon$$

Table 4.3 Explanation of Variables

__Yes	The percentage of “Yes” votes in the region
Freq	The total number of tweets in the region
PercentPositive	The percentage of the total tweets in the region scored “Positive”
PercentNegative	The percentage of the total tweets in the region scored “Negative”
Marijuana	Marijuana=1 if the legislation voted on in the region pertained to marijuana, Marijuana=0 if the legislation pertained to minimum wage
Urban	Urban=1 if the region is comprised of urban counties, Urban=0 if the region is comprised of rural counties

Table 4.4 Same Day Model

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Intercept	0.5918*	0.5678*	0.5880*	0.5758*	0.5542*	0.5810*	0.5617*	0.5746*	0.5602*	0.5287*	0.5572*	0.5304*
Total Volume	-0.0002					-0.0001			-0.0002			-0.0002
% Positive		0.0528					0.0478			0.0654		0.0557
% Negative			-0.0096					0.0064			-0.0374	0.0031
Marijuana				0.0163		0.0153	0.0117	0.0171				0.0091
Urban					0.0534***				0.0560**	0.0564**	0.0555**	0.0575***
Root MSE	0.0755	0.0748	0.0758	0.0754	0.0709	0.0765	0.0759	0.0767	0.0714	0.0705	0.0718	0.0741
R-Square	0.0076	0.0254	0.0005	0.0108	0.1259	0.0170	0.0308	0.0110	0.1437	0.1645	0.1327	0.1763
Adj R-Sq	-0.0266	-0.0082	-0.0340	-0.0233	0.0957	-0.0532	-0.0384	-0.0596	0.0826	0.1048	0.0707	0.0116

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

N: 997 tweets across 32 regions

Table 4.5 Day Before Model

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Intercept	0.5825*	0.5663*	0.5738*	0.5758*	0.5606*	0.5825*	0.5718*	0.5801*	0.5683*	0.5580*	0.5625*	0.5633*
Total Volume	-0.0032					-0.0032			-0.0039			-0.0040
% Positive		0.0262					0.0306			0.0203		0.0122
% Negative			-0.0188					-0.0256			-0.0155	0.0046
Marijuana				-0.0084		0.0000	-0.0135	-0.0118				0.0057
Urban					0.0215				0.0316	0.0181	0.0209	0.0309
Root MSE	0.0666	0.0681	0.0687	0.0687	0.0679	0.0686	0.0699	0.0705	0.0666	0.0696	0.0699	0.0736
R-Square	0.0647	0.0208	0.0047	0.0042	0.0271	0.0647	0.0309	0.0122	0.1198	0.0390	0.0303	0.1269
Adj R-Sq	0.0096	-0.0368	-0.0539	-0.0544	-0.0301	-0.0522	-0.0902	-0.1113	0.0097	-0.0812	-0.0909	-0.2089

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

N: 64 tweets across 20 regions

Table 4.6 Last Day Observed Model

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Intercept	0.6025*	0.5851*	0.5985*	0.5758*	0.5542*	0.5855*	0.5733*	0.5844*	0.5636*	0.5523*	0.5600*	0.5567*
Total Volume	-0.0045					-0.0046			-0.0054***			-0.00541***
% Positive		0.0238					0.0189			0.0109		-0.0118
% Negative			-0.0544					-0.0514			-0.0631	-0.0601
Marijuana				0.0232		0.0252	0.0191	0.0200				0.0219
Urban					0.0602**				0.0658**	0.0583**	0.0634**	0.0702**
Root MSE	0.0772	0.0789	0.0779	0.0788	0.0737	0.0775	0.0798	0.0786	0.0711	0.0748	0.0724	0.0717
R-Square	0.0601	0.0171	0.0433	0.0194	0.1427	0.0831	0.0296	0.0576	0.2285	0.1462	0.2006	0.2968
Adj R-Sq	0.0288	-0.0156	0.0114	-0.0133	0.1142	0.0198	-0.0374	-0.0074	0.1753	0.0873	0.1455	0.1616

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

N: 1,268 tweets across 32 regions

An OLS regression was utilized to predict the percentage of votes that were “yes” for each region. The OLS model was selected because the response variable, voting outcomes, are continuous and unique, ranging from 0.47737 to 0.742608. If the voting responses were transformed into discrete variables, with 1 meaning that an issue passed in that region and 0 meaning the issue failed, and another regression model was utilized, the response variable would cease being unique, as 29 of the 32 regions would bare a value of 1.

Three models were created to utilize tweet data from different time periods to predict the “yes” vote percentages. The Day Of Model used the observations occurring on election day, November 8th, 2016. The Day Before Model used the observations occurring on November 7th, 2016. The Last Day Observed Model used the observations from the day closest to preceding the election day that included an observation for a region. For twenty of the thirty two regions, the observations from November 7th were used. Some regions’ last tweet before the election dated as far back as October 21st.

The results suggest poor model fit, with the models failing to find a statistically significant relationship between media and voting outcomes. Compared to the Day Of Model, the Root MSE is reduced in all twelve versions of the Day Before Model. This could be due to Twitter users tweeting about their intended voting action the day prior to the election and reacting to the reported results on the evening of the election. Eleven of the twelve Last Day Observed models had higher Root MSE levels than the corresponding Day Of Model. The *Urban* variable was found to be significant at the .1 level or greater in all five versions in which it appeared in the Last Day Observed Model and the Day Of Model. However, the *Urban* variable was not found significant in any version of the Day Before Model. The Total Volume of tweets was found to be significant at the .1 level in versions 9 and 12 of the Last Day Observed Model.

Paired with low adjusted R-squared values, the limited occurrences of significant variables pertaining to Twitter activity suggest that the model is not an efficient tool for estimating the 2016 election results. The Urban variable routinely appeared in the model as a significant variable, but this characteristic is based on geographic information that is readily available, not Twitter specific data.

Limitations of the Data

Table 4.7 Day Of Summary Statistics

	FREQ	SumNeutral	SumNegative	SumPositive	SumNone
Mean	32.1613	18.1290	4.4516	9.5484	0.0323
Median	13	7	1	6	0
Min	1	0	0	0	0
Max	168	83	32	63	1

Table 4.8 Day Before Summary Statistics

	FREQ	SumNeutral	SumNegative	SumPositive	SumNone
Mean	3.3684	2.2105	0.5263	0.6316	0
Median	1	1	0	0	0
Min	1	0	0	0	0
Max	24	13	5	6	0

Table 4.9 Last Day Observed Summary Statistics

	FREQ	SumNeutral	SumNegative	SumPositive	SumNone
Mean	2.4063	1.5	0.375	0.5312	0
Median	1	1	0	0	0
Min	1	0	0	0	0
Max	24	13	5	6	0

The data’s most glaring limitation is the limited quantity of observations. Due to the one-year cap on accessing past data, only one election date could be studied. The observations that were able to be downloaded were greatly reduced through the process of narrowing results to those with a specific city location tag from within the state in which the legislation was voted on.

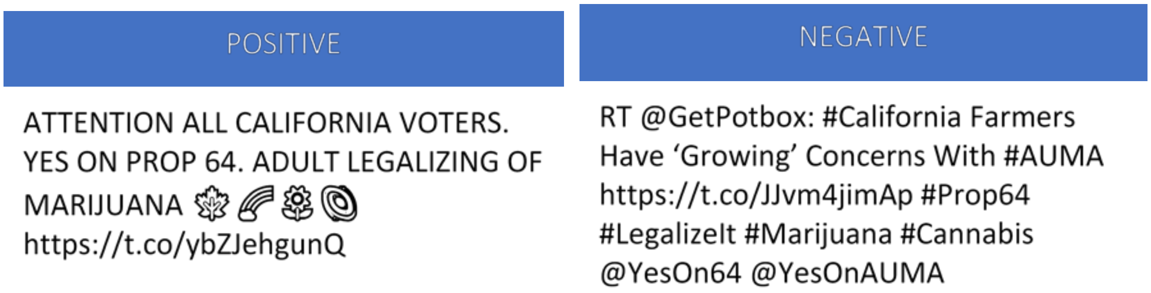
As a result, the mean number of total tweets in each region is a mere 32.16 tweets for the day of the election. Two regions relied on a single tweet to predict the percentage of yes votes for the Day Of Model. Only twenty of the thirty two regions produced even a single tweet the day before the election, with only one region producing more than 10 tweets.

The data from the day of the election was used due to the higher quantity of tweets that day than the days prior. While having additional data was beneficial, the model loses its predictive value to campaign teams. Predicting the results of an election multiple days in advance would allow campaigns to prioritize what geographic regions to focus efforts on in the final days before the election.

The problem of low observation quantity could be lessened by aggregating the volume of tweets over the two months preceding the election, as this data was accessed, rather than focus on specific day volume. However, this procedure would have resulted in extensive additional data cleaning and organization, as well as resulted in a loss of information about the trend of sentiment. The average sentiment of Twitter users from September 8th to November 8th may differ from the sentiment right before the election, which is the cumulating decision.

Relying on Sysomos' algorithm to score the sentiment of the tweets resulted in additional error. A tweet's status as either positive or negative, does not necessarily indicate the tweeters voting intention. The following tweets were scored oppositely, but both tweeters appear to support voting yes on Prop 64. Alternatively, manually scoring tweets to determine the sentiment of the content would be time consuming and tainted by human subjectivity.

Figure 4.1 Featured Prop 64 Tweets



POSITIVE	NEGATIVE
ATTENTION ALL CALIFORNIA VOTERS. YES ON PROP 64. ADULT LEGALIZING OF MARIJUANA 🍁🌈🌻🥑 https://t.co/ybZJehgunQ	RT @GetPotbox: #California Farmers Have ‘Growing’ Concerns With #AUMA https://t.co/JJvm4jimAp #Prop64 #LegalizeIt #Marijuana #Cannabis @YesOn64 @YesOnAUMA

Chapter 5 - Twitter as an Explanatory and Predictory Tool in the Stock Market

Models

Entered the Model:

$$(1)\ \ PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentTOne + \beta_3 PercentLexNex + \beta_4 PercentPos + \beta_5 PercentNeg + \beta_6 PTLagOne + \beta_7 PTLagTwo + \beta_8 PTLagThree + \beta_9 PTLagFour + \beta_{10} PTLagFive + \beta_{11} PLNLagOne + \beta_{12} PLNLagTwo + \beta_{13}\mathrm{PLNLagThree} + \beta_{14}\mathrm{PLNLagFour} + \beta_{15} PLNLagFive + \beta_{16} PosLagOne + \beta_{17} PosLagTwo + \beta_{18} PosLagThree + \beta_{19} PosLagFour + \beta_{20} PosLagFive + \beta_{21} NegLagOne + \beta_{22} NegLagTwo + \beta_{23} NegLagThree + \beta_{24} NegLagFour + \beta_{25} NegLagFive + \beta_{26} Hormel + \beta_{27} Campbell + \beta_{28} Sanderson + \varepsilon$$

Temporary Stacked Model:

$$(2)\ \ PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentLexNex + \beta_3 PercentTOne + \beta_4 PercentPos + \beta_5 PercentNeg + \beta_6 PLNLagFour + \beta_7 PLNLagFour + \beta_8 Sanderson + \varepsilon$$

Final Stacked Model:

$$(3)\ \ PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentLexNex + \beta_3 PercentPos + \beta_4 PercentNeg + \beta_5 PLNLagFour + \beta_6 NegLagOne + \beta_7 Sanderson + \varepsilon$$

Data

In this chapter, regression models were implemented to examine the percentage change of a specific food producing firm’s closing price from the previous day’s closing given changes in media volume. Historical stock prices were accessed through Yahoo! Finance. Data was collected for media pertaining to Tyson Foods (TSN), Hormel Foods Corporation (HRL), Campbell Soup Company (CPB), and Sanderson Farms, Inc. (SAFM) for trading days from

August 9th, 2016 to September 1st, 2017. The Twitter data collected from Sysomos utilized for these models was a frequency of the tweets posted per day that mentioned the firm in question and sentiment score frequencies, which tally the number of tweets for each sentiment category as scored by Sysomos' algorithm. The frequency of the Positive, Negative, Neutral, and None sentiment categories sum to the total frequency for the day. Frequency of news results per day was acquired using LexisNexis, through the Nexis Uni platform. A search for results using the same search terms and date range as the Twitter searches was ran and the results from the News category were downloaded. LexisNexis' news category includes newswires and press releases, industry trade press, newspapers, and web-based publications, among other news outlets. Binary variables were created to determine if the observation pertained to Sanderson, Campbell, Hormel, or Tyson Foods, with Tyson being the dropped variable in the models.

Procedure

A stepwise regression was used to identify the preferred model. In a stepwise regression, a minimum significance level for the p-values of the F-statistics for the specified variables is assigned (SAS). Variables are added to the model (1) one at a time and are only kept in the model if the p-value meets the selected standard. For this model, variables had to be significant at the 0.15 level or lower to be added. Once a variable is added, the model reexamines the other variables in the model to determine if the already added variables still meet the significance requirements. If any variable no longer meets the statistical significance standard, it is deleted before the next variable is considered. This method results in a final model, when all specified variables have been considered, leaving only variables that meet the significance requirements in the model and excluding those that do not (SAS, 2018).

In this study, the model determined in the final step of the stepwise regression was treated as a temporary model until multicollinearity issues could be resolved. The covariance of the variables in the model was determined and variables with a relationship with another variable with an absolute correlation value above .40 were examined to determine which to remove from the model. Each variable in question was individually ran in an OLS regression as the only independent variable, with *PercentClose* as the dependent variable. The variable with the higher resulting significance level was kept and the other was removed from the model. If both variables were significant at the same level, the variable in the regression that resulted in the higher R-squared value was kept. After the variables were removed to reduce multicollinearity, the stepwise regression was run again to produce the final model for each firm. If new variables were added to the resulting model after the removal of the problematic variables, the check for high correlation values was repeated. The temporary model became the final model if no relationships above an absolute .40 value were found. All observations contributing to the four final firm specific models were then merged to create a stacked data model.

While the step-wise regression technique was selected for this model, there are noted drawbacks to the method. While multicollinearity was addressed after the initial model had been run, the correlation between the variables can impact what variables are included in the temporary model. The method can also result in upward biased coefficients (Sribney, 1996). An alternative method to be considered in future research is the Lasso regression (Lease Absolute Shrinkage and Selection Operator). “Shrinkage” refers to shrinking data values towards a mean to combat high multicollinearity (Stephanie, 2015).

Table 5.1 Summary Statistics of All Candidate Variables in Stacked Data Model

Variable	Mean	Std Dev	Minimum	Maximum
PercentTOne	1.3626	2.1305	0.0599	42.8000

PTLagOneO	1.3626	2.1294	0.0599	42.8000
PTLagTwoO	1.3585	2.1275	0.0599	42.8000
PTLagThreeO	1.3703	2.1527	0.0599	42.8000
PTLagFourO	1.3688	2.1521	0.0599	42.8000
PTLagFiveO	1.3816	2.1739	0.0599	42.8000
PercentPos	1.9123	5.5358	0.0385	119.0000
PercentNeg	1.7536	3.2418	0.0345	46.0000
PosLagOne	1.9117	5.5359	0.0385	119.0000
PosLagTwo	1.9076	5.5352	0.0385	119.0000
PosLagThree	1.9340	5.5882	0.0385	119.0000
PosLagFour	1.9360	5.5886	0.0385	119.0000
PosLagFive	1.9486	5.5913	0.0385	119.0000
NegLagOne	1.7568	3.2448	0.0345	46.0000
NegLagTwo	1.7523	3.2442	0.0345	46.0000
NegLagThree	1.7595	3.2519	0.0345	46.0000
NegLagFour	1.7570	3.2518	0.0345	46.0000
NegLagFive	1.7741	3.2869	0.0345	46.0000
PercentClose	0.9999	0.0147	0.8551	1.0569
PercentLexNex	1.3401	1.5077	0.0769	18.2500
PLNLagOne	1.3403	1.5078	0.0769	18.2500
PLNLagTwo	1.3351	1.5011	0.0769	18.2500
PLNLagThree	1.3466	1.5458	0.0769	18.2500
PLNLagFour	1.3481	1.5477	0.0769	18.2500
PLNLagFive	1.3502	1.5469	0.0769	18.2500
PercentSP	1.0005	0.0052	0.9755	1.0222
Campbell	0.2500	0.4332	0.0000	1.0000
Sanderson	0.2500	0.4332	0.0000	1.0000
Hormel	0.2500	0.4332	0.0000	1.0000
Tyson	0.2500	0.4332	0.0000	1.0000

Table 5.2 Summary Statistics for Variables in the Final Campbell Soup Model

Variable	Mean	Std Dev	Minimum	Maximum
PercentSP	1.0005	0.005	0.9755	1.0222
PercentLexNex	1.2851	1.1770	0.1386	9.1250
PercentTOne	1.2775	1.2959	0.0599	11.8333

Note: 3,860 tweets and 7,160 LexisNexis articles were published during the observed period.

Table 5.3 Summary Statistics for Final Tyson Model

Variable	Mean	Std Dev	Minimum	Maximum
PercentSP	1.000	0.005	0.975	1.022
PercentTOne	1.514	3.280	0.184	42.800
NegLagFour	1.916	4.651	0.034	46.000
NegLagFive	1.918	4.650	0.034	46.000

Note: 4,091 tweets and 11,508 LexisNexis articles were published during the observed period.

Table 5.4 Summary Statistics for Final Hormel Model

Variable	Mean	Std Dev	Minimum	Maximum
PercentSP	1.000	0.005	0.975	1.022
PercentTOne	1.276	1.619	0.245	20.407
PercentPos	1.617	2.625	0.100	33.000
PTLagTwoO	1.275	1.619	0.245	20.407
PLNLagFive	1.425	1.782	0.077	16.000
PosLagFive	1.723	3.039	0.100	33.000
NegLagOne	1.848	3.384	0.045	36.000

Note: 8,689 tweets and 4,738 LexisNexis articles were published during the observed period.

Table 5.5 Summary Statistics for Final Sanderson Model

Variable	Mean	Std Dev	Minimum	Maximum
PercentSP	1.000	0.005	0.975	1.022
PercentLexNex	1.349	1.588	0.088	17.000
PercentPos	2.041	3.093	0.038	21.500
PercentNeg	1.813	2.646	0.095	25.000
PLNLagThree	1.352	1.590	0.088	17.000
PLNLagFour	1.354	1.590	0.088	17.000
PosLagFour	2.050	3.104	0.038	21.500
PosLagFive	2.076	3.117	0.038	21.500

Note: 8,973 tweets and 3,934 LexisNexis articles were published during the observed period.

PercentClose is the observed day’s closing stock price divided by the previous’ day’s closing stock price. All explanatory variables were converted into percentages as well, to reduce errors associated with the volume of media results varying across the firms. *PercentSP* is the closing stock price on the observed day for the S&P 500, divided by the previous day’s closing price. Due to some instances of the of the frequency of tweets for an observed day being zero, 1 was globally added to each tweet sentiment frequency (positive, negative, neutral, and none) and 4 was globally added to each day's total frequency count, to avoid losing observations due to error resulting from dividing by zero. *PercentTOne* is the upward adjusted total of tweets on the observed day divided by the previous day's total. *PercentPos* and *PercentNeg* were determined the same way, using the total number of positively and negatively, respectively, scored tweets. *PercentLexNex* is a measurement of the observed percentage of the day's total number of articles published compared to the prior day. *PLNLagOne*, *PLNLagTwo*, *PLNLagThree*, *PLNLagFour*, *PLNLagFive* represent the *PercentLexNex* value from one through five days previous. *TLagOne* through *TLagFive* follow the same procedure using the *PercentTOne* value. *PosLagOne* and *NegLagOne* through *PosLagFive* and *NegLagFive* track the lagged frequency for *PercentPos* and *PercentNeg*, respecitivley.

Table 5.6 Stock Price Model Results

	Tyson	**Hormel**	**Sanderson**	**Campbell Soup**	**Stacked Data**
Parameter	***Estimate***	***Estimate***	***Estimate***	***Estimate***	***Estimate***
Intercept	0.123	0.074	0.311	0.4198	0.468
PercentSP	0.879	0.930	0.685	0.5831	0.7550
PercentTOne	-0.002	-0.003			
PercentLexNex			-0.001	-0.0033	-0.0016
PercentPos		0.001	0.001		0.0001
PercentNeg			-0.001		-0.0008
PTLagOne					
PTLagTwo		-0.001			
PTLagThree					
PTLagFour					
PTLagFive					
PLNLagOne					
PLNLagTwo					
PLNLagThree			0.001		
PLNLagFour			0.002		0.0006
PLNLagFive		0.001			

PosLagOne					
PosLagTwo					
PosLagThree					
PosLagFour			0.001		
PosLagFive		-0.001	0.001		
NegLagOne		-0.001			-0.0002
NegLagTwo					
NegLagThree					
NegLagFour	0.000				
NegLagFive	0.000				
Hormel					
Campbell					
Sanderson					0.0028
Model Fit					
R-Squared	0.208	0.353	0.165	0.1696	0.1643
Adjusted R-Squared	0.195	0.335	0.138	0.1632	0.1587

Note: All estimates shown are significant at the 0.15 level or lower, per the step-wise regression

Interpreting the Stacked Data Model

The Stacked Model has an adjusted R-squared value of 0.1587, suggesting the media-driven model has limited ability to explain changes in stock prices for the firms in the sample. Three of the four individual firm models resulted in a higher adjusted r-squared value than the stacked model, implying that pooling observations from other firms in the industry does not aid in predicting a firm's stock price changes, but rather adds noise. The following figure illustrates some of the noise that arises from stacking the data. A spike in positive tweets occurred on September 27, 2016, with the positive tweets equaling 10800% of the previous day's count. All four firms saw positive tweets at least double from the previous day, with Hormel and Tyson seeing particularly large increases. Rather than the positive tweets resulting from an industry wide event, the tweets were specific to events within each firm. The Tyson spike was related to a "voluntary recall" of chicken nuggets. Tweets not mentioning "voluntary" typically did not classify as "positive", suggesting that a large quantity of tweets were classified as positive, despite being a negative event in the eyes of consumers. On the same day, Hormel announced a dividend and financial analysts predicted "earnings ahead" for the firm, with tweets responding to the news being largely classified as positive. This may provide further explanation for why the Tyson model did not find *PercentPos* to be a significant variable and the Hormel model found *PercentPos* to have a positive impact on the closing stock price change. This also likely partly explains why the model was able to explain less of the variance when observations from multiple firms were added.

Figure 5.1 Stacked Data Model Twitter Frequency

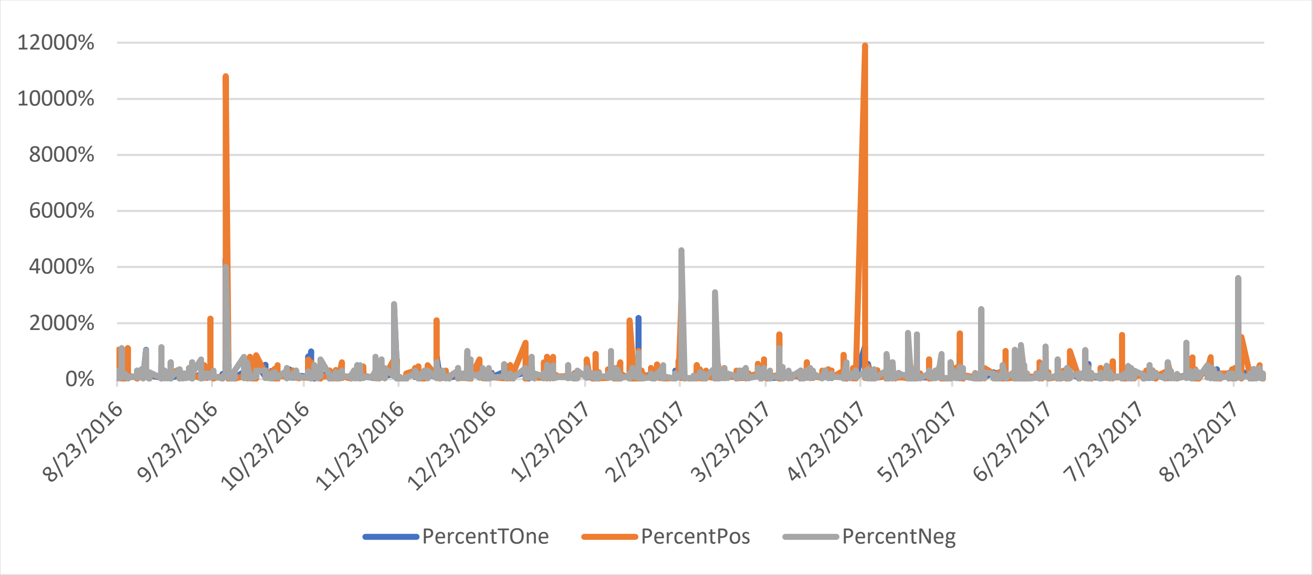


Figure 5.2 Tyson Model Twitter Frequency

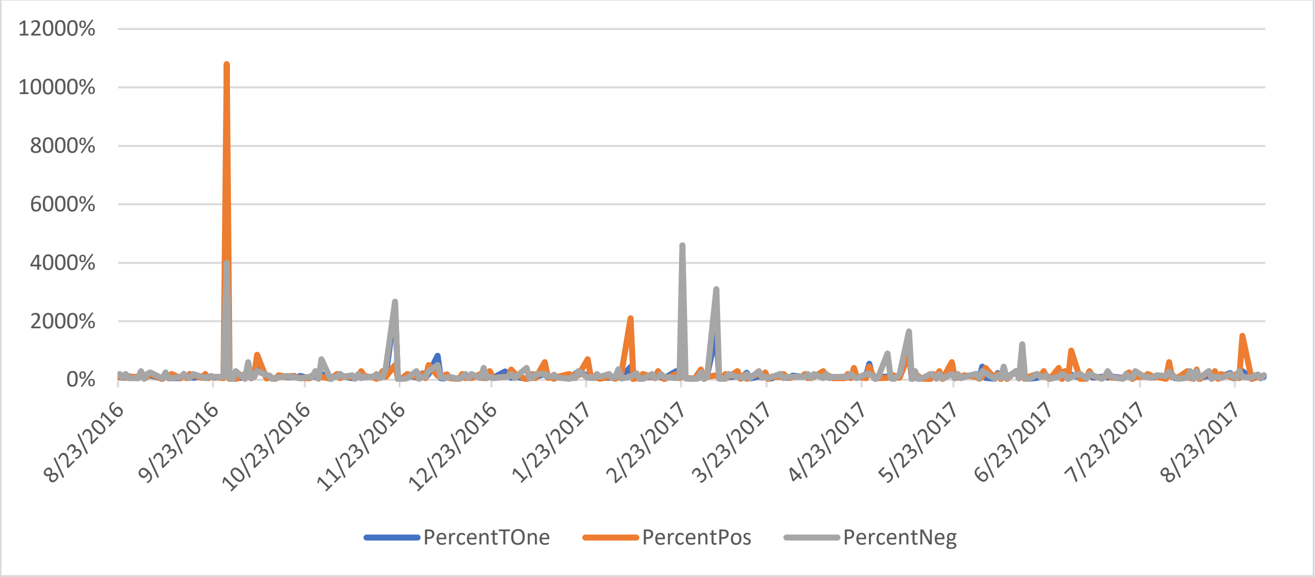


Figure 5.3 Hormel Model Twitter Frequency

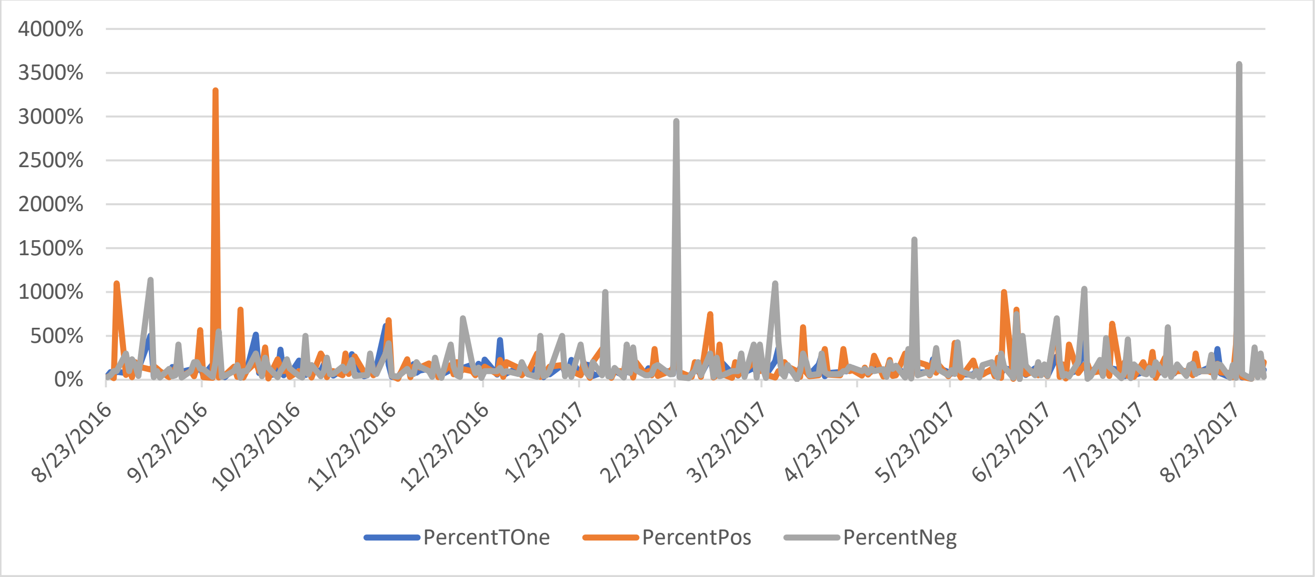


Figure 5.4 Sanderson Model Twitter Frequency

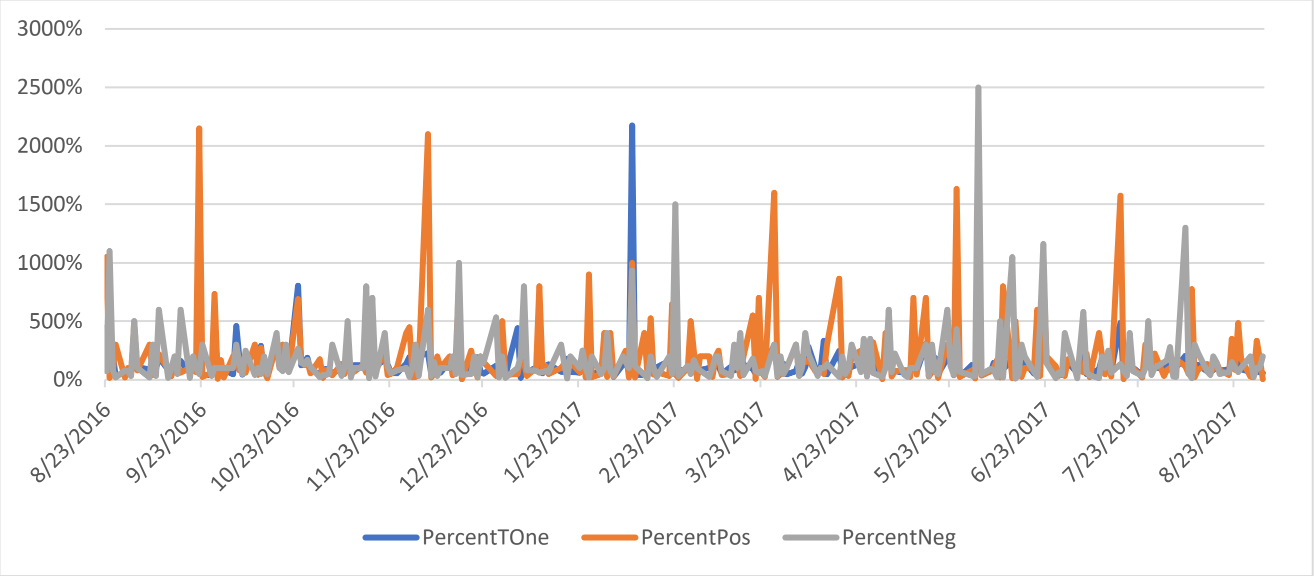
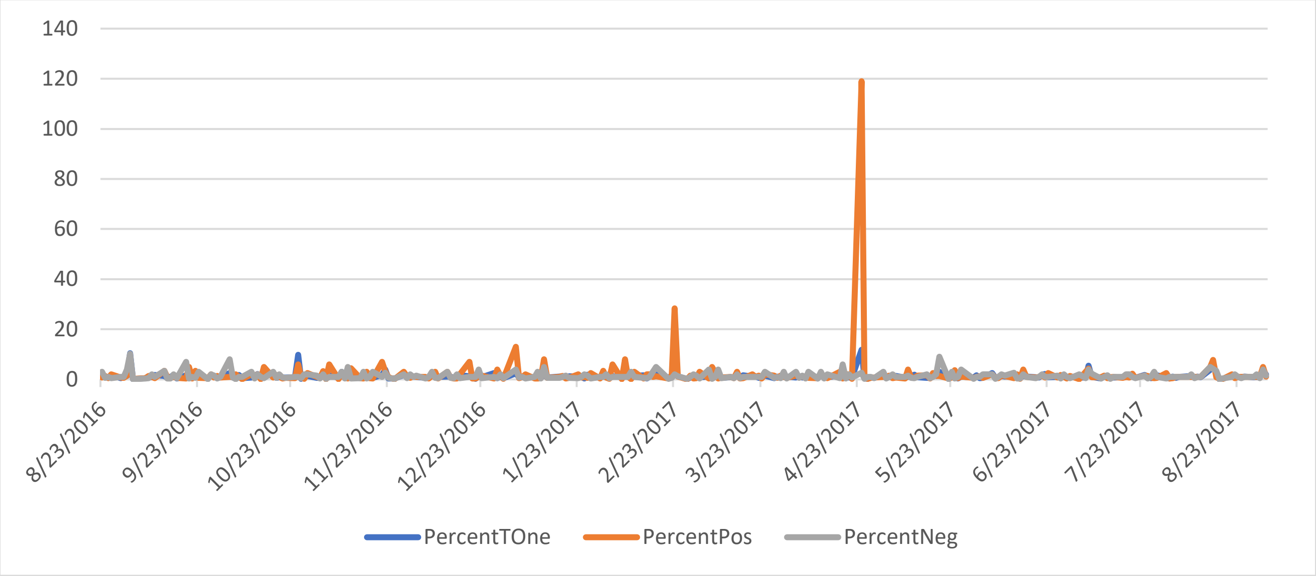


Figure 5.5 Campbell Soup Model Twitter Frequency



When the S&P closes 1% higher on the observed day than the previous day, the estimated stock is estimated to close .755% higher than the previous day. If the number of Lexis Nexis articles published increase by 1% over the previous day, the stock price is estimated to close 0.0016 % lower than the previous day, suggesting that LexisNexis coverage negatively impacts stock price. A 1% increase in LexisNexis percentage lagged four days implies a 0.0006% drop in stock price compared to the previous day, suggesting that LexisNexis coverage continues to have a negative impact on stock price several days after the increased media activity, but the magnitude of this impact decays over time. The *PercentPos* estimate implies that a 1% increase over the previous day's number of positive sentiment tweets results in a .0001% increase in stock price percentage. The *PosNeg* estimate implies a -0.0008% decrease in stock price percentage for a 1% increase in the percentage of negative tweets from the previous observation. These estimates follow the reasoning that positive sentiment tweets increase the valuation of a firm, while negative sentiment tweets decrease the value. The *NegLagOne* estimate of -0.0002 implies that negative sentiment tweets impact the stock price more the day after the increase in negative sentiment tweets than on the original day of increase. The Sanderson estimate implies that the observation pertaining to Sanderson results in a 0.0028% increase in stock price over the prior day, when compared to if the same observation values pertained to Tyson.

The *PercentSP* variable was found to be significant in all four individual firm models as well as the stacked model, with an estimate ranging from 0.5831 to 0.9300, implying that each firm moves in the same direction as the S&P 500. The variable was chosen to explain some of the trends and conditions of the stock market as a whole. Each time *PercentLexNex* or *PercentTOne*, the measures of LexisNexis and Twitter total volume, respectively, was found to be significant in a model, the coefficient was negative, implying that higher volumes of press have a negative impact on stock prices. Despite this each time a model found *PercentPos* to be significant, the estimate was positive, implying that while a general increase in tweets mentioning the firm negatively impact stock prices, positively scored tweets result in an increase in the stock price from the previous day.

Financial Implications

The fit of the model and the small values of the estimates imply there is little value to obtaining Twitter and LexisNexis data to explain the changes in stock prices on the average day. However, on days with large changes in media activity compared to the previous day, the financial value of the estimates is multiplied greatly. A media variable found to be significant was chosen from each individual firm model to study the impact on the day the maximum observation occurred, or the day when the variable in question increased the most in comparison to the previous day. The average and maximum financial values were determined by multiplying the estimated beta by the mean observation and the maximum observation. This value was then multiplied by the average stock closing price for the firm over the thirteen-month study period. For instance, *PercentTOne*, the observed day's total tweets as a percentage of the previous day's total, in the Tyson Model had an estimate of -0.002, implying when the total number of tweets increases 1% from the previous day, Tyson stock drops -0.002%. While the estimate is small, the mean observation is 1.514, meaning on average, the total quantity of tweets on an observed day is 154% of the previous day's total, resulting in an estimated stock price decrease of -0.0029%. Multiplied by Tyson's average stock closing price over the time period of $64.55, the *PercentTOne* variable is estimated on average, or a typical day, to account for a $0.19 decrease in closing stock price. Unless a stock holder owns a large number of Tyson stock, this value may not be worth the cost of accessing the data. This value increased greatly from the mean on September 27, 2016 when tweets pertaining to Tyson increased to 42.8 times higher than the previous day. The value of *PercentTOne* grew to an estimated decrease of $5.30 from the average closing stock price.

Table 5.7 Financial Impacts of Select Variables

	Hormel	**Tyson**	**Sanderson**	**Campbell Soup**
Variable:	*PercentPos*	*PercentTOne*	*PercentNeg*	*PercentLexNex*
Average Stock Closing:	$ 35.28	$ 64.55	$ 104.50	$ 56.68
Parameter Estimate:	0.001	-0.002	-0.001	-0.0033
Mean Observation:	1.617	1.514	1.813	1.2851
Maximum Observation:	33	42.8	25	9.125
Estimated Average Impact on Average Closing:	$ 0.03	$ (0.19)	$ (0.12)	$ (0.24)
Estimated Maximum Impact on Average Closing:	$ 0.71	$ (5.30)	$ (1.62)	$ (1.70)

Alternative Models

Additional versions of the stacked model were run to determine the explanatory value of the individual media sources, using the same process. The Everything Model is the original stacked model that includes information about Twitter sentiment frequencies, Twitter total frequency, LexisNexis frequency, the percent changes in the S&P 5500, and the firm specific binary variables, corrected for multicollinearity. The Twitter Sentiment model did not enter the *PercentTOne* variable or its associated lags. The Only Twitter Model excludes the *PercentLexNex* variable and its associated lags. The Only LexisNexis Model did not enter the *PercentTone*, *PercentPos*, or *PercentNeg* variable or their lags. The No Media Model only entered the *PercentSP* variable and the firm specific binary variables.

Twitter Sentiment Model

Entered the Model:

$$\begin{gathered}\mathit{PercentClose} = \beta_0 + \beta_1 \mathit{PercentSP} + \beta_2 \mathit{PercentLexNex} + \beta_3 \mathit{PercentPos} + \beta_4 \mathit{PercentNeg} + \\ \beta_5 \mathit{PLNLagOne} + \beta_6 \mathit{PLNLagTwo} + \beta_7 \text{PLNLagThree} + \beta_8 \text{PLNLagFour} + \beta_9 \mathit{PLNLagFive} + \\ \beta_{10} \mathit{PosLagOne} + \beta_{11} \mathit{PosLagTwo} + \beta_{12} \mathit{PosLagThree} + \beta_{13} \mathit{PosLagFour} + \beta_{14} \mathit{PosLagFive} \\ + \beta_{15} \mathit{NegLagOne} + \beta_{16} \mathit{NegLagTwo} + \beta_{17} \mathit{NegLagThree} + \beta_{18} \mathit{NegLagFour} + \beta_{19} \mathit{NegLagFive} + \\ \beta_{20} \mathit{Hormel} + \beta_{21} \mathit{Campbell} + \beta_{22} \mathit{Sanderson} + \varepsilon\end{gathered}$$

Final Twitter Sentiment Model:

$$\begin{gathered}\mathit{PercentClose} = \beta_0 + \beta_1 \mathit{PercentSP} + \beta_2 \mathit{PercentLexNex} + \beta_3 \mathit{PercentPos} + \beta_4 \mathit{PercentNeg} + \\ \beta_5 \text{PLNLagFour} + \beta_6 \mathit{NegLagOne} + \beta_7 \mathit{Sanderson} + \varepsilon\end{gathered}$$

Twitter Total Model

Entered the Model:

$$\begin{gathered}\mathit{PercentClose} = \beta_0 + \beta_1 \mathit{PercentSP} + \beta_2 \mathit{PercentTOne} + \beta_3 \mathit{PercentLexNex} + \beta_4 \mathit{PTLagOne} + \\ \beta_5 \mathit{PTLagTwo} + \beta_6 \mathit{PTLagThree} + \beta_7 \mathit{PTLagFour} + \beta_8 \mathit{PTLagFive} + \\ \beta_9 \mathit{PLNLagOne} + \beta_{10} \mathit{PLNLagTwo} + \beta_{11} \text{PLNLagThree} + \beta_{12} \text{PLNLagFour} + \beta_{13} \mathit{PLNLagFive} + \\ \beta_{14} \mathit{Hormel} + \beta_{15} \mathit{Campbell} + \beta_{16} \mathit{Sanderson} + \varepsilon\end{gathered}$$

Final Model:

$$\begin{gathered}\mathit{PercentClose} = \beta_0 + \beta_1 \mathit{PercentSP} + \beta_2 \mathit{PercentLexNex} + \beta_3 \mathit{PercentTOne} + \beta_4 \mathit{PLNLagOne} + \\ \beta_5 \text{PLNLagFour} + \beta_6 \mathit{Sanderson} + \varepsilon\end{gathered}$$

Only Twitter Model

Entered the Model:

$$\begin{gathered}\mathit{PercentClose} = \beta_0 + \beta_1 \mathit{PercentSP} + \beta_2 \mathit{PercentTOne} + \beta_3 \mathit{PercentPos} + \beta_4 \mathit{PercentNeg} + \\ \beta_5 \mathit{PTLagOne} + \beta_6 \mathit{PTLagTwo} + \beta_7 \mathit{PTLagThree} + \beta_8 \mathit{PTLagFour} + \beta_9 \mathit{PTLagFive} + \\ \beta_{10} \mathit{PosLagOne} + \beta_{11} \mathit{PosLagTwo} + \beta_{12} \mathit{PosLagThree} + \beta_{13} \mathit{PosLagFour} + \beta_{14} \mathit{PosLagFive}\end{gathered}$$

$$+\beta_{15} NegLagOne + \beta_{16} NegLagTwo + \beta_{17} NegLagThree + \beta_{18} NegLagFour + \beta_{19} NegLagFive + \beta_{20} Hormel + \beta_{21} Campbell + \beta_{22} Sanderson + \varepsilon$$

Final Model:

$$PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentPos + \beta_3 PercentNeg + \beta_4 \text{NegLagOne} + \beta_5 NegLagFour + \beta_6 Sanderson + \varepsilon$$

<u>Only LexisNexis Model</u>

Entered the Model:

$$PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentLexNex + \beta_3 PLNLagOne + \beta_4 PLNLagTwo + \beta_5 \text{PLNLagThree} + \beta_6 \text{PLNLagFour} + \beta_7 PLNLagFive + \beta_8 Hormel + \beta_9 Campbell + \beta_{10} Sanderson + \varepsilon$$

Final Model:

$$PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 PercentLexNex + \beta_3 PLNLagOne + \beta_4 \text{PLNLagFour} + \beta_5 Sanderson + \varepsilon$$

<u>No Media Model</u>

Entered the Model:

$$PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 Hormel + \beta_3 Campbell + \beta_4 Sanderson + \varepsilon$$

Final Model:

$$PercentClose = \beta_0 + \beta_1 PercentSP + \beta_2 Sanderson + \varepsilon$$

Table 5.8 Stock Market Model Fit

Model	R-Square	Adjusted R-Square
Everything Model	0.1643	0.1587
Twitter Sentiment (no total)	0.1643	0.1587
Twitter Total (no sentiment)	0.1656	0.1607
Only Twitter (LexisNexis excluded)	0.1393	0.1343
Only LexisNexis (Twitter excluded)	0.1386	0.1344
No Media (only PercentSP and dummy variables)	0.0801	0.0783

Table 5.9 Stock Market Model Error

Model	Root MSE	Error Change
Everything Model	0.0135	-4.67%
Twitter Sentiment	0.0135	-4.67%
Twitter Total	0.0135	-4.75%
Only Twitter	0.0137	-3.14%
Only LexisNexis	0.0137	-3.14%
No Media	0.0141	

If Root Mean Square Error, a measure of the standard deviation between the estimated values and observed values, is used as the primary determinant of fit, the model that excludes all variables pertaining to media frequency is the poorest fit, implying that adding information about media frequency aids in reducing the unexplained variance. The adjusted R-Square values also imply this. The RMSE is identical for the Twitter Only and LexisNexis Only models, suggesting that the two media sources are nearly identical in value added to predicting the percentage change in stock price. The RMSE is reduced further and the adjusted R-Square is increased when both LexisNexis and Twitter are considered together, implying that while the two media sources explain a similar amount of the variation, there is a slight difference in the part of the variation that is being explained. The model that enters variables pertaining to Twitter total tweets, Twitter sentiment, LexisNexis, S&P 500 closing changes, and what firm the observation pertains to (Everything Model) is identical to the model that does not consider total tweets due to the PercentTOne model being removed from the temporary Everything Model due to high correlation with the PercentPos and PercentNeg variables.

Out-of-Sample Application

The stacked model was used to predict the percentage change in stock prices for two additional food producing publicly traded firms to determine the ability of the model to estimate out of sample data. ConAgra (ticker symbol CAG) and Pilgrim's Pride (ticker symbol PPC) were selected due to their similarities to the four firms in the stacked model and the availability to data. Twitter data was collected from Sysomos and LexisNexis articles were accessed for the two additional firms over the same time period as the stacked model firms. The following tables show the Root Mean Square Error for the models estimated using the stacked model as well as the change in error compared to running the model that includes no variables pertaining to media coverage. The Only Twitter Pilgrim's Pride model was the only model to see a reduction in errors from the base No Media model. This suggests that using the changes in media volumes of other firms to predict out of sample data is not efficient, but rather adds noise to the model.

Table 5.10 ConAgra

	Root MSE	Error Change
Everything Model	0.01260	5.34%
Twitter Total	0.01238	3.66%
Only Twitter	0.012647	5.69%
Only LexisNexis	0.012195	2.20%
No Media	0.011927	

Table 5.11 Pilgrim’s Pride

	Root MSE	**Error Change**
Everything Model	0.01807	2.09%
Twitter Total	0.01837	3.66%
Only Twitter	0.01764	-0.32%
Only LexisNexis	0.01853	4.51%
No Media	0.01769	

Chapter 6 - Twitter's Exchangeability with Traditional Print Media

On March 5, 2017, the USDA announced the presence of HPAI in a commercial chicken flock in Tennessee (USDA, 2017). The impacted flock consisted of 73,500 birds. In 2017, the average liveweight of a U.S. broiler was 6.18 pounds and for the week ending on March 6, the average price was 87 cents per pound (National Chicken Council, 2011) (USDA, 2017). By these averages, the birds in the compromised flock had a market value over 3.95 million dollars. Within days, 27 countries and the European Union temporarily banned imports of poultry from Lincoln County, Tennessee, the site of the outbreak, with several of the countries instituting bans from broader regions within the U.S., likely negatively impacting the profits of some disease-free poultry operations (AgNet West, 2017). Over the following twenty days, an additional case of HPAI in a poultry flock was discovered in Tennessee, as well as occurrences of low pathogenicity avian influenza (LPAI) in Alabama, Kentucky, and Georgia (USDA, 2017). Over the period for which Twitter information was accessed for this study, additional events occurred that impacted animal health and welfare, including the confirmation of a bovine spongiform encephalopathy (BSE) case in Alabama and a USDA announcement of changes designed to end horse soring, all potentially having economic impacts on players involved in the animal production chain (USDA, 2017).

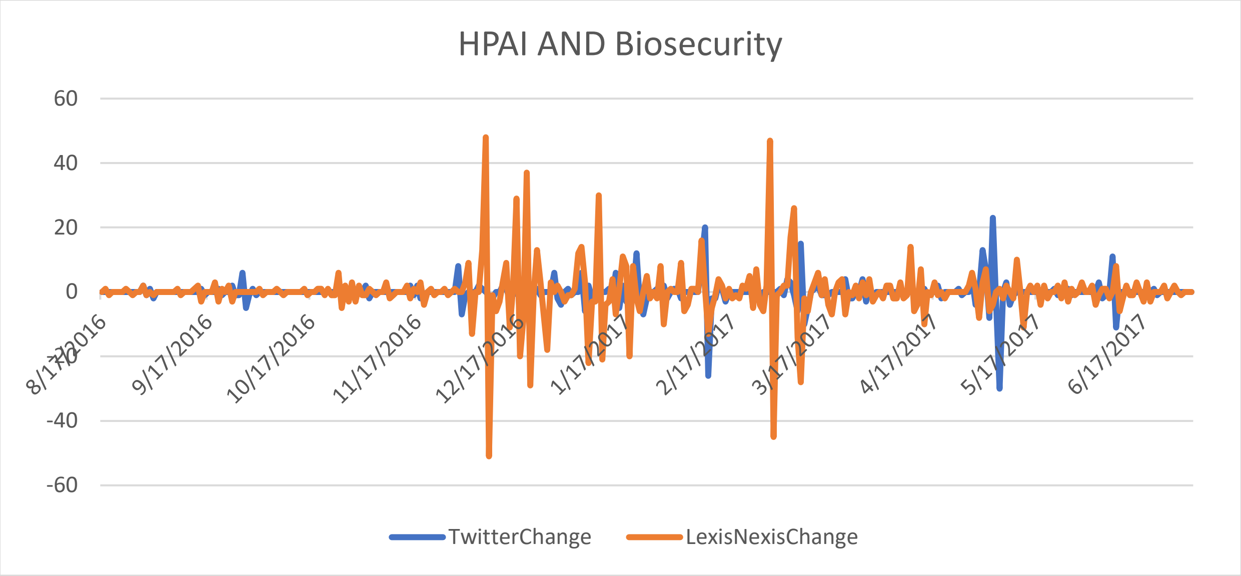
Due to the potential financial implications of events that impact animal health and welfare, there is an incentive to monitor consumer sentiment about such events. If players in the animal production industry are able to gain insight into consumer reactions, there is potential to craft response plans that target the specific concerns of consumers. However, if the cost to obtain data on consumer sentiment is high, any benefits from increased information may be negated.

For this study, information pertaining to traditional media coverage of topics was collected from LexisNexis and Twitter information was collected from Sysomos. Nexis Uni, the academic version of LexisNexis, is a standard inclusion in many college’s database subscription portfolios. According to LexisNexis, over 9 million students and faculty have access to the search platform (LexisNexis Academic, 2018). Alternatively, Sysomos caters primarily to corporations and typically require minimum contracts of one year, which were quoted for this study to cost over $30,000 (Sysomos, 2018). A special two month contract was negotiated for this study at a cost of $7,278). In June of 2016, Sysomos introduced the “Sysomos in the Classroom” program that allows college professors to apply for the opportunity to incorporate Sysomos created tools into their curriculum, but at this time there is no academic version of the program that would allow students and faculty at a participating university to freely access the Sysomos search platform (Sysomos, 2018).

At the time of data collection this study, Sysomos offered one of the most extensive past histories of Twitter data of any major social media analytic firm with 100% of the volume from the preceding thirteen months (Sysomos, 2018). The volume of tweets compared to posts from other social media platforms makes archiving the past activity expensive. Twitter specific packages for the statistical software R are being continually updated to expand the possibilities associated with extracting Twitter data through a financially free avenue, including programs that perform sentiment analyses on the tweets. At the time tweets were downloaded in 2017, the most popular tweet extraction methods using R could only access the preceding week to the day of the download, due to Twitter’s public API limiting the availability of historical tweets (Twitter, 2018).

Due to the costs of Twitter data access, there is an incentive to test the exchangeability of traditional print media data for Twitter data in studies attempting to capture information about market or consumer behavior through media. To compare the two sources for this study, a comparison was completed using media articles and tweets relating to animal health. Results pertaining to porcine epidemic diarrhea virus (PEDV) and highly pathogenic avian influenza (HPAI) were downloaded, as well as results pertaining to the respective topics and biosecurity.

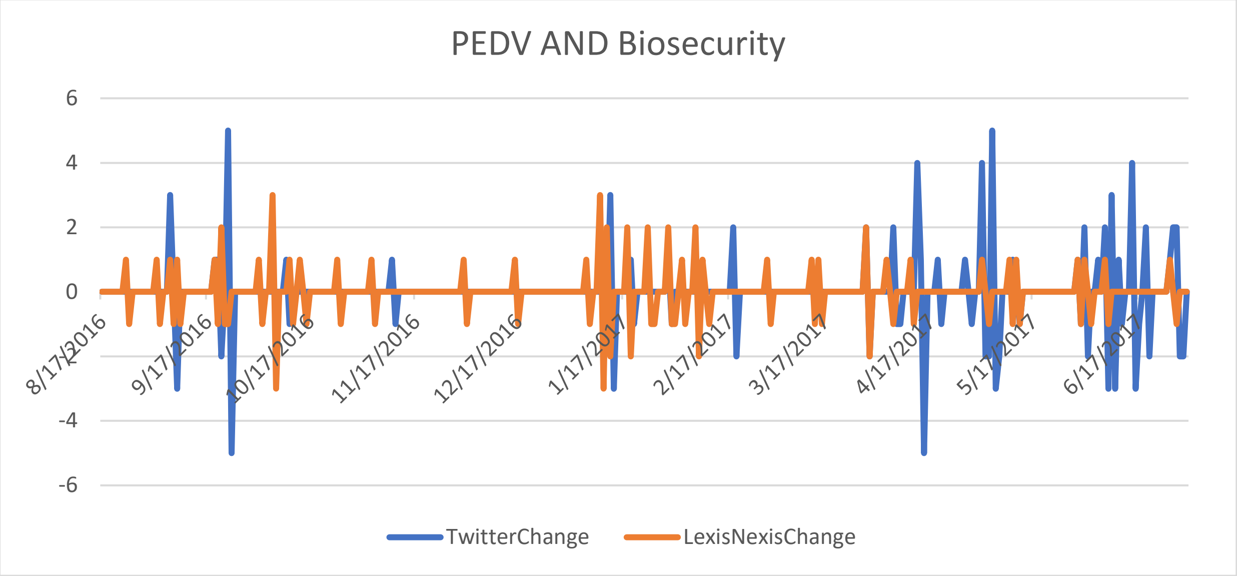
Figure 6.1 HPAI AND Biosecurity Frequency Change



Note: Corr=0.06571

Variable	Mean	Std Dev	Minimum	Maximum
TwitterChange	0.0000	3.7876	-30	23
LexisNexisChange	0.0031	8.2250	-51	48

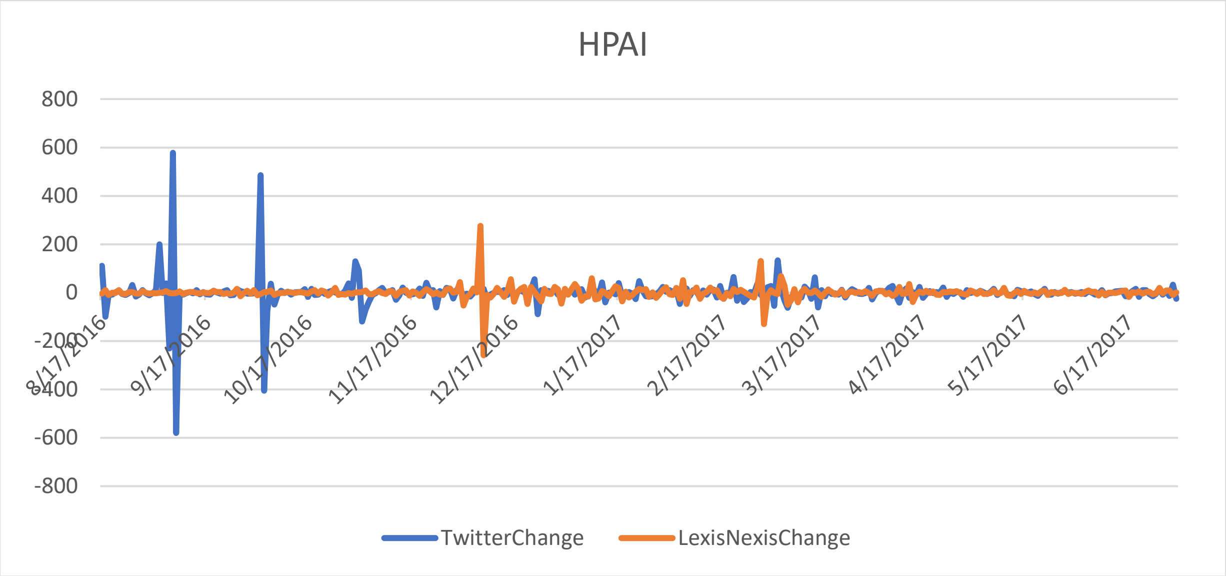
Figure 6.2 PEDV AND Biosecurity Frequency Change



Note: Corr= 0.00806

Variable	Mean	Std Dev	Minimum	Maximum
TwitterChange	0.0031	1.0171	-5	5
LexisNexisChange	0.0000	0.6635	-3	3

Figure 6.3 HPAI Frequency Change

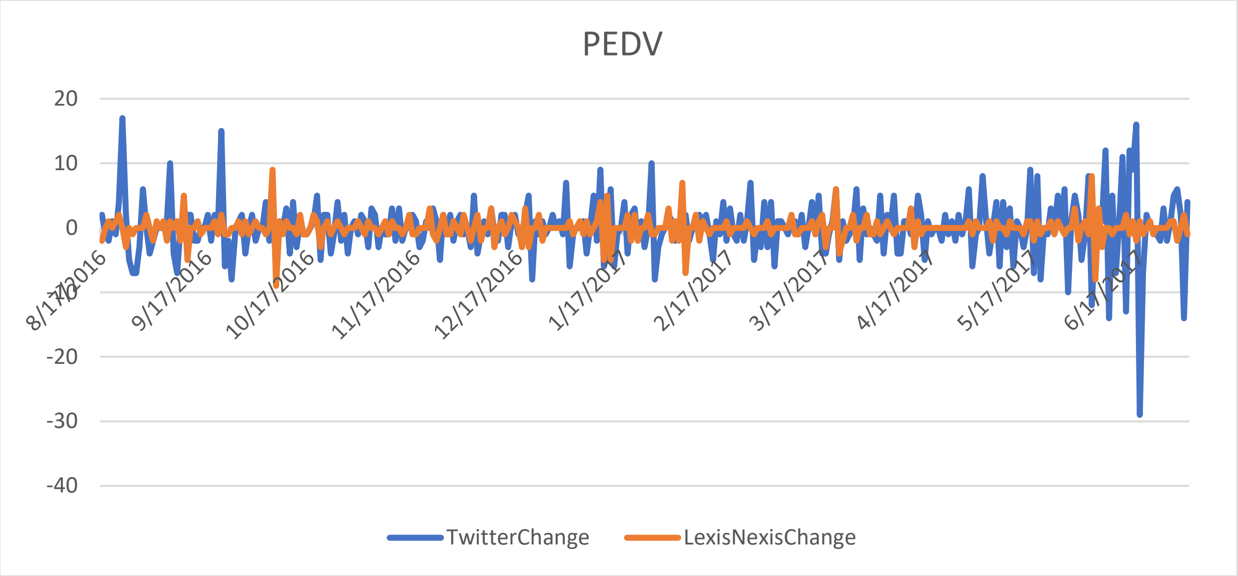


Note: Corr= 0.02529

Variable	Mean	Std Dev	Minimum	Maximum
TwitterChange	-0.0125	65.2185	-580	578

LexisNexisChange	0.0627	27.7541	-259	276

Figure 6.4 PEDV Frequency Change



Note: Corr= 0.06571

Variable	Mean	Std Dev	Minimum	Maximum
TwitterChange	0.0125	4.4333	-29	17
LexisNexisChange	-0.0031	1.7618	-9	9

Models

Media Explanatory Models

$$(1)\ \ TwitterChange = \beta_0 + \beta_1 TLagOne + \beta_2 TLagTwo + \beta_3 TLagThree + \beta_4 TLagFour + \beta_5 TLagFive + \varepsilon$$

$$(2)\ \ TwitterChange = \beta_0 + \beta_1 LNLagOne + \beta_2 LNLagTwo + \beta_3 LNLagThree + \beta_4 LNLagFour + \beta_5 LNLagFive + \varepsilon$$

$$(3)\ TwitterChange = \beta_0 + \beta_1 TLagOne + \beta_2 TLagTwo + \beta_3 TLagThree + \beta_4 TLagFour + \beta_5 TLagFive + \beta_6 LNLagOne + \beta_7 LNLagTwo + \beta_8 LNLagThree + \beta_9 LNLagFour + \beta_{10} LNLagFive + \varepsilon$$

$$(4)\ \ LexisNexisChange = \beta_0 + \beta_1 LNLagOne + \beta_2 LNLagTwo + \beta_3 LNLagThree + \beta_4 LNLagFour + \beta_5 LNLagFive + \varepsilon$$

$$(5)\quad LexisNexisChange = \beta_0 + \beta_1 TLagOne + \beta_2 TLagTwo + \beta_3 TLagThree + \beta_4 TLagFour + \beta_5 TLagFive + \varepsilon$$

$$(6)\quad LexisNexisChange = \beta_0 + \beta_1 TLagOne + \beta_2 TLagTwo + \beta_3 TLagThree + \beta_4 TLagFour + \beta_5 TLagFive + \beta_6 LNLagOne + \beta_7 LNLagTwo + \beta_8 LNLagThree + \beta_9 LNLagFour + \beta_{10} LNLagFive + \varepsilon$$

TwitterChange is the total frequency of tweets containing the search phrase (HPAI, HPAI AND biosecurity, PEDV, PEDV AND biosecurity) on the day prior to the observed day, subtracted from the observed day's total frequency. *LexisNexisChange* follows the same formula, using news articles published on LexisNexis rather than tweets. *TLagOne*, *TLagTwo, TLagThree*, *TLagFour,* and *TLagFive* are the *TwitterChange* values lagged one through five days. *LNLagOne*, *LNLagTwo*, *LNLagThree*, *LNLagFour*, *LNLagFive* are the *LexisNexisChange* values lagged one through five days.

A simple OLS regression (1) was run to estimate the change in the sum of tweets from the previous day to the observed day using the tweet change variable lagged for the succeeding five days for the time period from August 17, 2016 to July 1, 2017. The model was then run again using the lagged change in LexisNexis articles for the succeeding five days as the explanatory variable rather than the lagged tweet changes (2), with the objective of determining the ability of LexisNexis data to explain Twitter observations, and thus identifying the exchangeability of the two media sources. A step-wise OLS regression was then run to determine the ability of the five lagged days of Twitter and LexisNexis frequency changes together to explain the change in tweets from the previous observed day. The step-wise regression entered each proposed variable into the model individually, only keeping variables found to be significant at the 0.15 level or lower in the model. The process was then repeated using the change in LexisNexis articles as the dependent variable (4), (5), and (6).

Results

Table 6.1 Results of Twitter Explaining Twitter

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	-0.0662	0.0224	0.0000	0.0194
TLagOne	-0.6656*	-0.4215*	-0.3389*	-0.5962*
TLagTwo	-0.3679*	-0.3029*	-0.3711*	-0.5371*
TLagThree	-0.2056*	-0.3463*	-0.1332**	-0.3985*
TLagFour	-0.0945	-0.1975*	-0.1109***	-0.2541*
TLagFive	-0.1024***	-0.0792	-0.1634*	-0.2398*
R-Square	0.3187	0.1902	0.1805	0.3037
Adj R-Square	0.3073	0.1773	0.1674	0.2926

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

Table 6.2 Results of LexisNexis Explaining Twitter

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	0.0259	0.0125	0.0005	0.0031
LNLagOne	-0.0884	0.4038**	-0.0022	0.1372
LNLagTwo	-0.1382	0.2794	-0.0274	0.2035
LNLagThree	-0.1452	0.0423	0.0073	0.1140
LNLagFour	-0.2227	0.2035	-0.0527	0.0347
LNLagFive	-0.1070	-0.2857	-0.0104	-0.0176
R-Square	0.0053	0.0417	0.0148	0.0089
Adj R-Square	-0.0106	0.0263	-0.0009	-0.0069

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

Table 6.3 Twitter and LexisNexis Explaining Twitter

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	-0.0643	0.0424	0.0000	0.0194
TLagOne	-0.6587	-0.4156	-0.3389	-0.5962
TLagTwo	-0.3410	-0.3108	-0.3711	-0.5371
TLagThree	-0.1517	-0.3609	-0.1332	-0.3985
TLagFour		-0.2062	-0.1109	-0.2541
TLagFive		-0.0859	-0.1634	-0.2398
LNLagOne				
LNLagTwo				
LNLagThree				
LNLagFour		0.3292		
LNLagFive				
R-Square	0.3104	0.2071	0.1805	0.3037
Adj R-Square	0.3039	0.1918	0.1674	0.2926

Note: All estimates shown are significant at the 0.15 level or lower, per the step-wise regression.

Table 6.4 Results of LexisNexis Explaining LexisNexis

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	0.2064	-0.0042	0.0112	0.0022
LNLagOne	-0.6581*	-0.8368*	-0.5220*	-0.8119*
LNLagTwo	-0.5494*	-0.6144*	-0.5244*	-0.6296*
LNLagThree	-0.4789*	-0.5044*	-0.3769*	-0.5457*
LNLagFour	-0.4048*	-0.2469*	-0.3066*	-0.3926*
LNLagFive	-0.3499*	-0.1248**	-0.2646*	-0.3085*
R-Square	0.3638	0.4334	0.2840	0.4398
Adj R-Square	0.3536	0.4243	0.2726	0.4308

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

Table 6.5 Results of Twitter Explaining LexisNexis

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	0.0606	-0.0006	0.0031	0.0016
TLagOne	0.0121	-0.0055	0.1012	-0.01529

TLagTwo	-0.0077	-0.0396	-0.2592***	-0.0468
TLagThree	-0.0177	-0.0216	0.0021	-0.0609
TLagFour	-0.0174	-0.0032	-0.1905	-0.0341
TLagFive	-0.0202	0.0086	-0.0365	0.0516
R-Square	0.0030	0.0099	0.0207	0.0157
Adj R-Square	-0.0129	-0.0059	0.0050	0.0000

Note: Statistical significance is indicated with asterisks to the right of the estimate. The 1%, 5%, and 10% levels are represented by *, **, and ***, respectively.

Table 6.6 Twitter and LexisNexis Explaining LexisNexis

	HPAI	**PEDV**	**HPAI and Biosecurity**	**PEDV and Biosecurity**
Parameter	*Estimate*	*Estimate*	*Estimate*	*Estimate*
Intercept	0.2064	-0.0004	0.0112	0.0036
TLagOne		0.0360	0.1839	
TLagTwo				
TLagThree				-0.0501
TLagFour				-0.0684
TLagFive				
LNLagOne	-0.6581	-0.8483	-0.5276	-0.8195
LNLagTwo	-0.5494	-0.6396	-0.5270	-0.6304
LNLagThree	-0.4789	-0.5253	-0.3749	-0.5393
LNLagFour	-0.4048	-0.2593	-0.3103	-0.3828
LNLagFive	-0.3499	-0.1421	-0.2569	-0.2986
R-Square	0.3638	0.4412	0.2911	0.4502
Adj R-Square	0.3536	0.4305	0.2774	0.4378

Note: All estimates shown are significant at the 0.15 level or lower, per the step-wise regression.

The R-Square values for the model suggest that lagged Twitter and LexisNexis frequency change observations explain some of the variation in the observed day's frequency's change from the previous day. The R-Square values also suggest that lagged Twitter and LexisNexis frequency change observations explain very little of the change in the other media source's observed day frequency, implying a lack of exchangeability in using the two sources for models featuring media frequency related variables. Across the four animal health topics, the mean R-

Square values for LexisNexis explaining LexisNexis and Twitter explaining Twitter are 0.3803 and 0.2483, respectively, implying lagged LexisNexis data explains LexisNexis frequency changes better than lagged Twitter data explains Twitter frequency changes. This could suggest that Twitter users respond to events reported in the news that impassion them, causing drastic spikes in tweets, while ignoring smaller events that trigger the news publication, but not consumer response. LexisNexis explains Twitter models have a mean R-Square value of 0.0177 and Twitter explains LexisNexis models have a mean R-Square value of 0.0123, implying lagged LexisNexis data has a slight advantage over lagged Twitter in explaining the contrasting media source's frequency changes.

While Twitter data was found to be a unique source of information from LexisNexis data, significant commercial value cannot be implied. Knowing that tweets contain information about consumer sentiment and activity is a far cry from being able to extract specific, usable data to be used in predictive and explanatory modeling. Identifying and executing methods for cleaning the data is likely a limiting factor for many firms and organizations due to the sheer cost.

Across the four subjects in the Twitter Explaining Twitter model, the 18 out of the 20 lagged Twitter variables were found to significant at the 10% level or lower. Only one out of the 20 variables were found to significant at the 10% level or lower for the LexisNexis Explaining Twitter model. In the LexisNexis Explaining LexisNexis model, all twenty lagged LexisNexis variables were found significant at the 5% level or lower, with only one lagged variable found significant in the Twitter Explaining LexisNexis model.

Figure 6.5 Absolute Twitter Decay

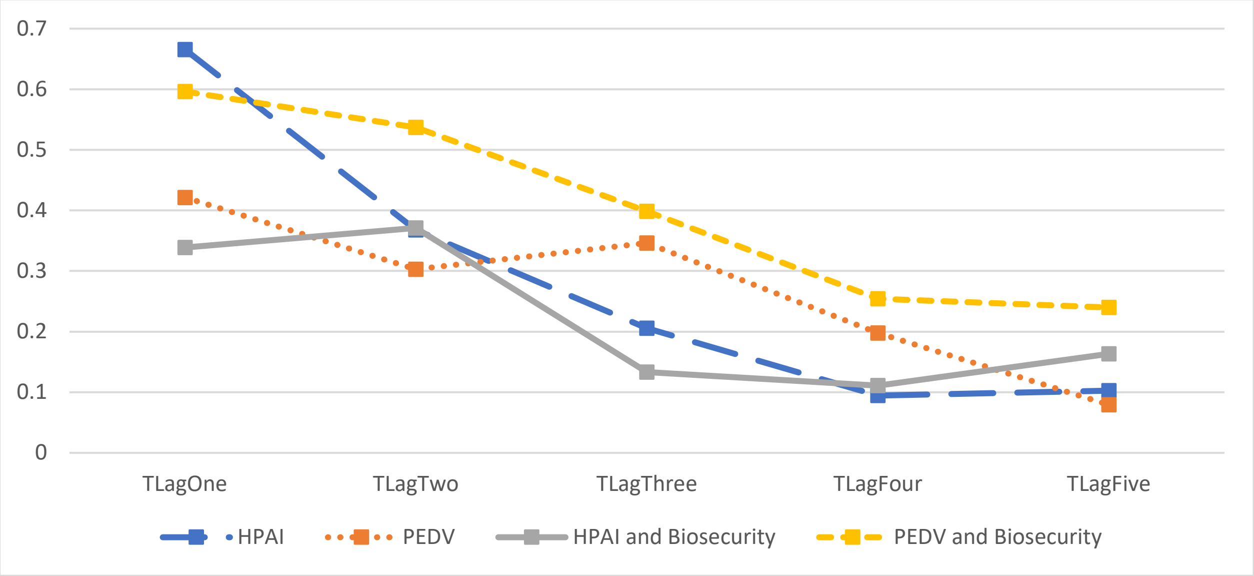
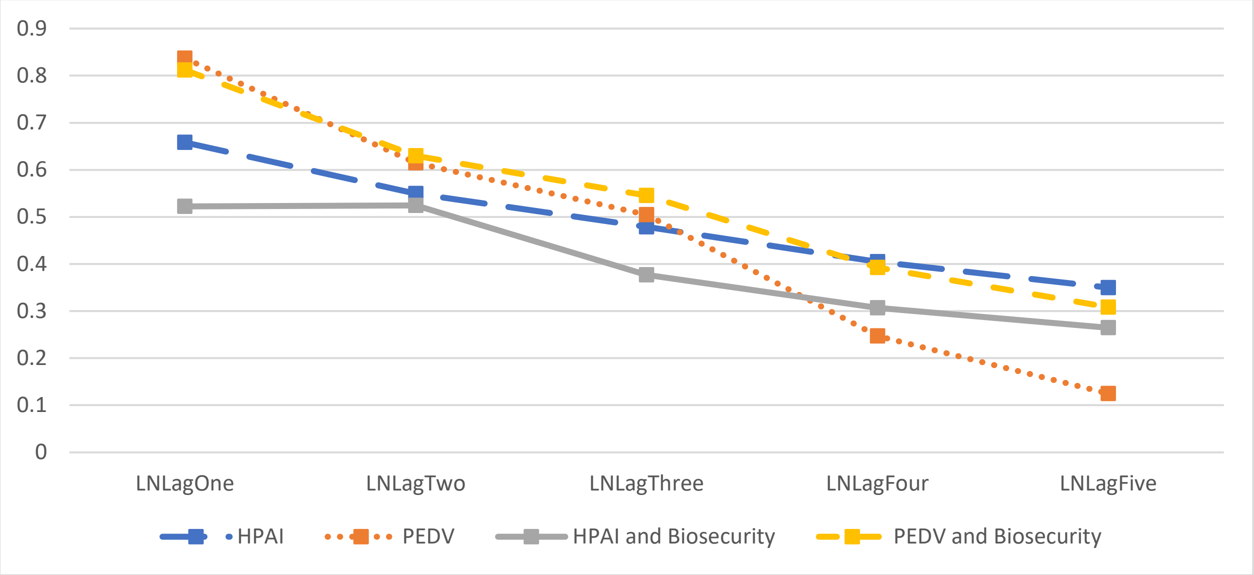


Figure 6.6 Absolute LexisNexis Decay



Figures 6.5 and 6.6 show the decaying impact the lagged variables have on the observed day's change in frequency, using the estimates from the Twitter Explaining Twitter model and LexisNexis Explaining LexisNexis model, respectively. The patterns show a mostly declining impact over the course of the five lagged days, although the estimates do not reach zero nor become insignificant by the fifth lagged day. Additional lagged variables could be added to the

model to determine the point at which the lagged observations no longer have an impact on the current observation.

Due to the lack of sentiment scoring for LexisNexis models, it is not possible to use this data to determine if the sentiment of the observations of the two sources are similar. A frequency observation also does not shed light onto the subject of the media results, potentially useful information when determining the exchangeability of the two sources.

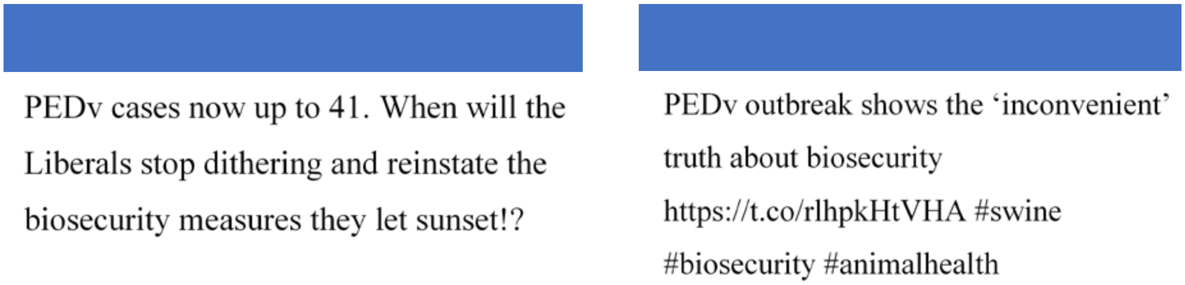
Table 6.7 Root Mean Square Error (RMSE)

	Dependent Variable		
	LexisNexis (6)	***Twitter (3)***	***Error Change***
HPAI	22.3142	54.4149	144%
HPAI and Biosecurity	6.9916	3.4561	51%
PEDV	1.3296	3.9855	200%
PEDV and Biosecurity	0.4975	0.8555	72%

Error change was determined by calculating the percentage change in RMSE for Model 3 compared to Model 6. For HPAI and PEDV, the RMSE was higher for the models in which Twitter was the dependent variable. When “biosecurity” was added to the search terms, the Twitter RMSE was lower than the LexisNexis RMSE, although the Twitter adjusted R-square values were lower for both biosecurity versions than the LexisNexis adjusted R-square values.

For tweets referencing PEDV and biosecurity, 86.1% feature a hyperlink. Hyperlinks often lead to a news article about the topic, sometimes shared by the author’s or news organization’s account, sometimes by a consumer passing along information he or she found interesting to their friends and family. 82% of HPAI and biosecurity tweets, 84.6% of PEDV tweets, and 86.3% of HPAI tweets included a hyperlink. Tweets featuring hyperlinks are less likely to feature the Twitter user’s views or opinions than tweets without. This is demonstrated in the featured tweets below. Despite this high occurrence of hyperlinks, change in LexisNexis results was not found to be an effective instrument to explain changes in Twitter activity.

Figure 6.7 Featured PEDV Tweets



PEDv cases now up to 41. When will the Liberals stop dithering and reinstate the biosecurity measures they let sunset!?

PEDv outbreak shows the 'inconvenient' truth about biosecurity https://t.co/rlhpkHtVHA #swine #biosecurity #animalhealth

Chapter 7 - Conclusion

Over the span of a decade, Twitter went from creation to the publication of 500 million tweets each day (Politico Staff, 2016). This study sought to identify uses for Twitter data as a research tool in the agricultural economics field and analyze the value added by utilizing the data. In Chapter 4, this study failed to find a significant relationship between Twitter coverage and voting outcomes pertaining to marijuana and minimum wage in the 2016 election. The low number of tweets meeting the search and location requirements likely weakened the explanatory value of the variables. Additionally, the explanatory models used tweets that were published too close to the election date to be valuable to campaign teams as a predictive tool.

In Chapter 5, this study found little value in using tweet volume changes to explain food producing firm stock price changes on the average day. But did find a significant relationship on peak days of activity changes, such as on days when food recalls or firm announcements occurred. The percent change in tweets from the previous day to the observed day was found to explain up to $5.30 of the change in price of Tyson stock on the day of maximum change. A stacked data model performed poorly in predicting the stock price changes of food producing firms outside of the sample, suggesting that firm specific events and stock market wide trends, as tracked by the S&P 500, better explained changes in closing stock prices than the activity of other firms in the food producing sector. To receive maximum financial benefit from using Twitter data, an investor must have continual access to Twitter data to make decisions throughout the day.

In Chapter 6, Twitter and LexisNexis were not found to be strong substitutes for each other when used to track the coverage of animal health events, implying that Twitter provides unique information from traditional print media. For both traditional print media and tweets, the number of articles published on the five preceding days were found to have a lasting, although decreasingly so, impact on the number of articles published on the observed day, implying that media coverage of animal health events lasts beyond the day of the event and can continue to influence consumers and investors.

Much like Twitter's unpredictable rise to prominence, the next innovation in data source will likely develop quickly and in an unexpected way. Researchers need to be prepared to find ways to test the economic value of these new sources and determine their efficiency. Commercialization of data may keep some sources out of reach for academic researchers and too expensive to provide financially beneficial information to firms and investors. Understanding the value that the data provides will allow purchasers of the data to make more economical decisions.

Further Research

Throughout this study, the Twitter data's biggest limiting factor was the constraint to only a single year of tweets. Continual access to the data would open the door to a number of opportunities to further this study, including the opportunity to test the explanatory models' ability to predict the next year of stock market prices or election results. Additional years of historical data would provide a better base from which to build these models.

Opportunities exist for more fully transforming and utilizing the available data. In this study, tweets were assigned a binary sentiment score. Using an algorithm to assign a polarity score from -1 to +1 would provide further information about the influence of the sentiment of a tweet. Creating Twitter-sourced variables, beyond frequency and sentiment, such as the reach and region of a tweet, would provide additional information that would potentially provide clarity into what tweets have an impact on consumer behavior. Location tagged data was used to narrow data sets to the state of interest for the election models, but could have been utilized in the stock price and animal health models to determine if particular regions held stronger influence than others. Sysomos provided “authority” score could be used as a variable to determine if the expertise or prominence of the tweet author impacted stock prices.

There is potential for Twitter data to be used to explain consumer decisions in the food industry. Tweets mined for purchase intent after food related events, such as the announcement of an *E. coli* outbreak in ground beef, a lettuce recall, or the introduction of a new genetically modified product, could be utilized in models explaining consumption or sales.

References

AgNet West. (2017, March 6). *USDA Confirms H7 Avian Influenza in Tennessee*. Retrieved from AgNet West: http://agnetwest.com/78592/

Azar, P. D., & Lo, A. W. (2016). The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds. *The Journal of Portfolio Management*, 123-134.

Batrinca, B., & Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI & Society*, 89-116.

Elmer E. Rasmuson Library. (2016, October 12). *Boolean Searching*. Retrieved from Elmer E. Rasmuson Library: https://library.uaf.edu/ls101-boolean

Kanakaraj, M., & Guddeti, R. M. (2015). NLP based sentiment analysis on Twitter data using ensemble classifiers . Chennai, India: IEEE.

LexisNexsis Academic. (2018). *LexisNexis Academic-General Information*. Retrieved from LexisNexis Academic: https://www.lexisnexis.com/communities/academic/w/wiki/30.lexisnexis-academic-general-information.aspx

Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2014). Predicting movie Box-office revenues by exploiting large-scale social media content . *Multimedia Tools and Applications*, 1509-1528.

National Chicken Council. (2017, September 26). *U.S. Broiler Performance*. Retrieved from National Chicken Council: https://www.nationalchickencouncil.org/about-the-industry/statistics/u-s-broiler-performance/

Newberry, C. (2016, August 11). *Top Twitter Demographics That Matter to Social Media Marketers*. Retrieved from Hoot Suite: https://blog.hootsuite.com/twitter-demographics/

Politico Staff. (2016, November 8). *Election results 2016 by state and county* . Retrieved from Politico: https://www.politico.com/story/2016/10/election-results-2016-by-state-and-county-229735

Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLoS ONE*, 1-21.

SAS. (n.d.). *Model-Selection Models*. Retrieved from SAS/STAT(R) 9.2 User's Guide, Second Edition: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect030.htm

Sribney, B. (1996). *What are some of the problems with stepwise regression?* . Retrieved from Stata: https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/

Stephanie. (2015, September 24). *Lasso Regression*. Retrieved from Statistics How To: http://www.statisticshowto.com/lasso-regression/

Sysomos. (2018). *Sysomos*. Retrieved from Sysomos: https://sysomos.com/about/

Tonsor, G. T., Mintert, J. R., & Schroeder, T. C. (2010). U.S. Meat Demand: Household Dynamics. *Journal of Agricultural and Resource Economics*.

Twitter. (2018). *Choosing a historical API*. Retrieved from Twitter Developer: https://developer.twitter.com/en/docs/tutorials/choosing-historical-api

USDA. (2017). *Livestock, Dairy, and Poultry Outlook.* USDA.

USDA. (2017, March 15). *Livestock, Dairy, and Poultry Outlook*. Retrieved from USDA: https://www.ers.usda.gov/webdocs/publications/82850/ldp-m-273.pdf?v=42809

USDA. (2017, March 6). *USDA Confirms Highly Pathogenic H7 Avian Influenza in a Commercial Flock in Lincoln County, Tennessee*. Retrieved from USDA Animal and Plant Health Inspection Service : https://www.aphis.usda.gov/aphis/newsroom/news/sa_by_date/sa-2017/hpai-tn