Acknowledgements

First and foremost, I would like to thank my Thesis Advisor, Dr. Amar Das. Your mentorship transformed my college academic experience and inspired me to pursue data science in healthcare. I am forever grateful for the opportunity and guidance you provided over the past two years. I am also indebted to Steven Andrews for his contribution to the project. And Rebecca Faill, thank you for technical expertise, always being available when I have questions, and most of all your support.

Next, I would like to extend my gratitude to Professor Michael Herron. I have appreciated our weekly check-ins, the discipline you have helped me find with my work, and your continuous support. I would also like to thank Jordana Composto for taking the time to give me detailed feedback and being a part of this journey with me.

Thank you to the Dataminr Data Analysis team for the opportunity to learn and grow in ways that built a foundation for this project. Thank you to Dr. Shilpa Murthy for long conversations about breast cancer screening management, my first research experience, your mentorship, and most of all your friendship.

I would like to thank my friends and Anna for making me step away when I needed a break. Thank you Ava for being my on-call editor and cheerleader.

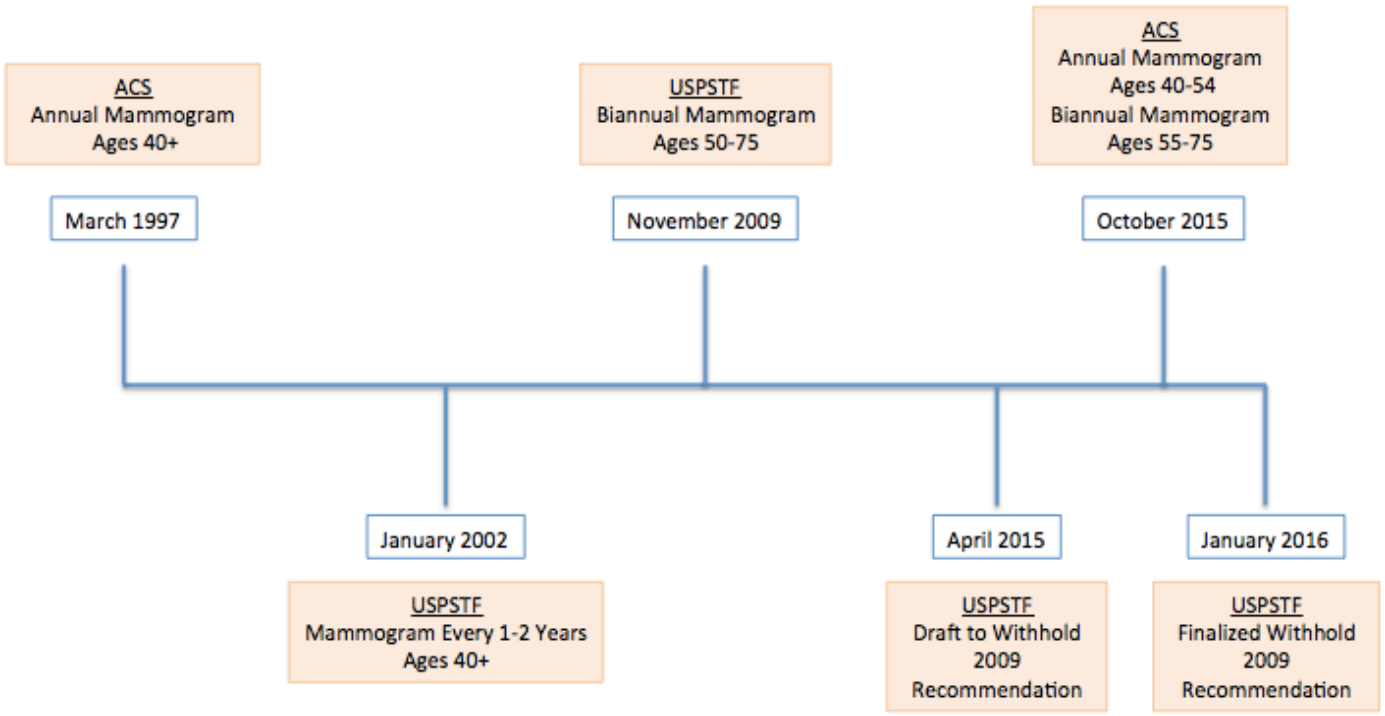
Finally, thank you to my grandparents who suffered the tragedy of breast cancer, you were in my heart through every read tweet. And thank you most to my parents.

Table of Contents

Introduction

In recent years, breast cancer screening recommendations have stimulated controversy between key health care actors. The timeline below, "Overview of Recommendation History," shows how breast-screening recommendations have changed over the past 20 years. Current controversy stems from the 2009 United States Preventative Services Task Force (USPSTF) announcement. In November 2009, the USPSTF changed its 2002 recommendations such that women should start biennial screening at age 50 instead of 40 [29]. Meanwhile, medical societies, such as the American Cancer Society (ACS), continued to recommend that women begin annual screening at age 40 [28]. Since 2009, the USPSTF responded to criticism of

Overview of Recommendation History



[28,29]

the 2009 guideline change, by pledging to investigate the benefits and harms of screening before the age of 50. In April 2015, the USPSTF released a draft recommendation that would renew their 2009 recommendation based on recent research. The Task Force finalized this recommendation in January 2016. In October 2015, the ACS responded to the lack of evidence supporting screening before age 50 and changed its recommendations to begin annual screening at age 45 and biannual screening at age 55.

The inconsistency of recommendations has spurred confusion for women seeking breast cancer care management and prevention [25]. In order to respond appropriately to patients' concerns, healthcare providers and policy makers can benefit from understanding the level of public confusion and anxiety regarding the conflicting recommendations. One solution for public sentiment and opinion surveillance is using social media as a data source. This research study uses Twitter to analyze public response to the use of breast cancer screening technologies. We use Twitter as our social media outlet due to several features that make it an attractive application for research. The research methodology consists of data-mining tweets related to breast cancer screening from the Twitter Search application programming interface (API) in order to analyze public conversation pertaining to the controversial health technology. The study will use textual analytics and machine-learning methods to perform an automated sentiment analysis of tweets from April 16, 2015 to April 15, 2016, a year in which both the USPSTF and the ACS changed its recommendations. The analysis will provide both a

qualitative and quantitative understanding of trends in Twitter sentiment as well as tweet type concerning the promotion or avoidance of screening technologies as recommendations change.

The results of this study will be distributed into three chapters: In Chapter 1, I examine the value of Twitter data for surveillance of consumer health opinion, particular with breast cancer screening. In Chapter 2, I present the methods and results of predicting sentiment and tweet category using machine-learning techniques. In Chapter 3, I determine the changes in content and volume of tweets over time and assess how changes in breast cancer screening recommendations affect the Twitter dialogue.

Chapter 1[1]

[1] The following research was collected through a key word search in the Dartmouth College Summon

Introduction: Why Twitter as a Data Source?

Opinions are a central piece of data in researching human behavior [7]. Businesses, organizations, and policy makers, who seek to understand how a population feels about a certain topic traditionally, relied on data in the form of surveys or opinion polls. This method requires active participation of studied groups, is costly, and produces limited data. However, the Internet has enabled new methods of opinion-mining that address these limitations; social media data does not directly require active participants, is inexpensive to use, and produces large data sets [7].

In recent years, the increasing use of social media and micro-blogging[2] has introduced a rich source of information containing public and personal opinions. Social media creates unprecedented spaces for information sharing and social network formation. The most popular micro-blogging services are Facebook, Tumblr, and Twitter [18]. These services allow users from around the world to connect about their experiences, share their opinions, and discuss current events. Thus, the information provided on social media sites has incredible potential for opinion research.

The most common micro-blogging service for data mining is Twitter. Twitter allows a user to post texts of maximum 140-characters that can be immediately seen by their followers. Twitter has an average of 307 million monthly users, creating a large volume of daily tweets that are available to the public through the Twitter

[2] The activity or practice of making short, frequent posts.

API[16]. The value of Twitter to data mining is not just in its large scale and availability, but also the content that it provides: 13% of online adults[3] use Twitter, most users tweeting daily and on their phones [16]. These users are likely to tweet about a broad range of issues that affect their lives, providing an abundance of diverse information regarding personal experiences.

Opinion-mining through Twitter analysis remains a difficult task; Twitter data are noisy and those who tweet may not be representative of the non-Twitter population. When tweets are collected through a keyword search, the data often contains spammers and a certain percentage of non-relevant tweets. Meda et al. addressed this concern by suggesting machine-learning methods to detect spammers in Twitter datasets. Meda et al. support the claim that machine-learning[4] is the backbone necessary for analyzing Twitter data [17]. The methods suggested by Meda et al. will be used in this research study to determine whether tweets are relevant to breast cancer screening.

Additionally, the extent to which Twitter can accurately represent the non-Twitter population is unknown. Papers such as *Modeling the Impact of Lifestyle on Health at Scale* address this limitation. Sadilek et al. hypothesized that patterns of disease can be detected by applying machine-learning algorithms to Twitter data. The authors validated their results by comparing their findings to traditional method of health

[3] In 2015, 84% of American adults use the Internet [26].

[4] Machine-learning is the practice of developing algorithms that can be trained to make predictions on data.

surveys. This study showed that their Twitter analysis was consistent with previous epidemiological work. Thus, the authors suggested that though Twitter users may differ from non-Twitter users, Twitter analysis still provided information that can benefit the entire population. The study also demonstrates that the usefulness of Twitter may differ across research topics. [16]

Sentiment Analysis of Twitter

Twitter has a wide variety of information that is dispersed between its users; tweets range from government public statements to individuals' personal thoughts. Therefore, analyzing opinions through Twitter streams requires thorough investigation of the nature of language pertaining to the researched topic. Machine-learning techniques for opinion-mining[5] aim to minimize false predictions due to language barriers such as sarcasm, misspelling, or slang words. Additionally, the abundance of information provided by the Twitter API is almost impossible to analyze. Thus, researchers agree that automated sentiment analysis methods are best for monitoring opinions in Twitter streams.

In 2010, Bifet et al. determined methods for sentiment analysis of tweets. Bifet et al. proposed that analyzing tweets requires a methodology that enables scientists to cope with large amount of data, which is the biggest obstacle to Twitter research.

[5] Opinion mining (also known as sentiment analysis) determines the attitude towards a particular topic. The sentiment of language is generally classified as being positive (the user feels positive about the topic), negative (the user feels negative about the topic), or neutral (the user does not express emotion.

Additionally, Bifet et al. raise concerns about a secondary challenge to Twitter analysis, hidden variation in Twitter language. Bifet et al. explain that machine-learning techniques can minimize challenges such as sarcasm and irony. They hypothesized that Twitter analysis requires building machine-learning classifiers with a subset of data, known as the training data. Then, using the classifiers to detect trends in the entire data set. This process enables researchers to study only a subset of tweets to get results from a large data set. The authors concluded that Twitter could potentially provide real-time analysis of a particular event or topic and advocated for further research in the field. [1]

Kumar et al. explain that the value of Twitter for sentiment analysis is "unmatched" due to its variance in user type and global location; Twitter allows users to build both local and global connection networks in a timely and easy way. Kumar et al. motivation was to expand on previous research that applied sentiment analysis to Twitter in order to determine public opinion on political issues. Kumar et al. performed a case study on gauging public mood on Twitter through sentiment analysis. The authors collected Twitter data and then determined the sentiment of each tweet through a manual scoring system. They extracted opinion words, a combination of adjectives, and verbs and adverbs in a subset of tweets and then manually performed sentiment scoring based on predetermined classes. The authors chose to classify a tweet's sentiment as either positive, negative, or neutral. The conclusion of this study were that Twitter suggests promising results for the

field of sentiment analysis and further research should be performed to investigate best techniques for Twitter as as data set. [2]

Barbosa et al. researched sentiment detection techniques from Twitter and were particularly concerned with extracting information from "noisy" data. The authors used Wekaito Environment for Knowledge Analysis (Weka), a machine-learning tool that builds classifiers for automated sentiment detection. The authors explained methods for preprocessing the data, such as building the appropriate n-gram list in Weka for classification. Additionally, their methodology consisted of a labeling process for the test set to input to Weka for building the models. [3]

Weka has gained popularity for building automated classification in the literature. Umadevi's paper *Sentiment Analysis Using Weka* (2014) describes a methodology for sentiment analysis using Weka. As Barbosa et al. explain, Umadevi defines automated sentiment analysis as a supervised learning task that requires each text in a training set to have a predetermined class label. Using Weka, Umadevi tockenized tweets into individual words using the *StringtoWordVector*[6] application. He also used stemmers to convert words with the same root word such as "*drives," "driving," "driver," "drive,"* to a single word. Additionally, Umadevi used the preprocessing application that eliminates stopwords. Stopwords are words that are used so frequently that they do not contribute value to the sentiment of the text. When building the classifiers, Umadevi used three-fold cross validation in Weka to

[6] A function in Weka used to preprocess the data in order to build classifiers.

predict how accurately the model could predict classes of a test set. The accuracy of his predictions ranged from 87- 92% correctly classified, depending on the classifier and preprocessing technique. The methodology in this paper also consisted of testing different preprocessing techniques and comparing results. This step is necessary for each specific set of data due to the unique language pertaining to each domain. The conclusion of this paper was that the technique produced relatively good results. This research project will use much of the methodology suggested by Umadevi. [4]

Applications of Sentiment Analysis of Twitter

In 2013, Neethu et al. published a paper that recommended machine-learning techniques for sentiment analysis of Twitter data to determine opinions about products. They chose to perform a case study on tweets that mentioned certain electronic products, such as laptops and mobile devices. Neethu et al. claim that when someone wants to purchase or use a product, they review it online in posts and blogs. However, the online database for opinions regarding an electronic item are often too large to analyze. Thus, there is a need for an automated methodology. The authors discovered sentiment pertaining to an electronic product by data-mining a set of tweets based off keywords, preprocessing the tweet for analysis, create feature vector by building n-grams and assigning scores -1, 0, 1 to the vector for sentiment classification. The authors then used a Matlab simulation to determine which classifiers performed best on the training set. The results showed that the feature vector technique performs well when evaluating electronic products. [5]

Pak et al. agree that manufacturing companies may be interested in using Twitter as a corpus for opinion-mining and sentiment analysis. These authors explain that, because people use micro-blogging to express opinions about different topics, Twitter data could reveal the extent to which people are positive about a topic. Pak et al. investigated techniques that improved the results of sentiment classification and found that n-gram tokenization worked best in this application [6]. I will build on these results and use n-gram tokenization in my research analysis as well.

Bing Liu discusses other applications of sentiment analysis in his book *Sentiment Analysis and Opinion-mining*. Liu shows that, in 2010, a surge in Twitter sentiment analysis occurred due to an explosive growth of social media: Tumasjan analyzed Twitter sentiment to predict election results; Bollen et al. predicted the stock market based off Twitter sentiment; and, Liu predicted box-office revenues by data-mining tweets related to the film. Liu found that his predictions were quite accurate. Liu's studies reveal that sentiment analysis has reached almost every possible domain, from consumer products to services, political elections and healthcare [7]. My research project will contribute to the application of Twitter sentiment analysis in healthcare.

Healthcare as an Application of Twitter Sentiment Analysis

The healthcare industry is transforming towards a system in which the provider considers patients' opinion about their care. In recent years, the development of

"patient-centered healthcare" has gained momentum in the progress of health services. During the same decade healthcare began to experience these changes, social media emerged as a medium for information and opinion sharing. Social media has also proven to shape events in societies such as the integral role Twitter and Facebook had in the Arab Spring, when a video of a Tunisian fruit vendor setting himself on fire sparked protests through social media [28].

Rozemblum et al. describe the parallel transformation of healthcare and social media as the "perfect storm." Rozemblum et al. determines that patients are increasingly engaged in the quality of their healthcare and share their opinions and experiences via social media. Rozemblum et al. suggest that key players in healthcare should use the information provided through social media to acquire patient feedback, which was previously determined by paper-based surveys. The authors also believe that social media analysis in the healthcare domain will not have too many biases. [8]

One of the first studies which uses natural language processing and sentiment analysis was used to determine patient feedback, was Greaves et al. *Use of Sentiment Analysis for Capturing Patient Experiences From Free-Text Comments Posted Online*. Greaves et al. hypothesized that analytical techniques such as sentiment analysis would provide information to improve the quality of healthcare. The authors used machine-learning techniques to detect how patients felt about their quality of care. The study was conducted to predict whether a patient would recommend their

hospital, whether the hospital is clean, and whether the patient felt as if they were treated well. The authors explained that the "free-text" format of an online comment more accurately portrays the sentiment of a patient than a traditional survey, which asks the patient to score their care in a quantitative rating. Greaves et al. collected 6412 online comments about hospitals from the English National Health Service website and used Weka to perform a sentiment analysis. The authors then compared the findings of their machine-learning classification to the findings of a paper-based survey. The results showed that the automated predictions matched the results of the survey. The authors concluded that their findings suggest that social media could have major value for improving the quality of healthcare and that their methodology can be used in other applications of machine-learning. [9]

Several other papers have investigated the use of opinion-mining in the healthcare domain. Chew and Eysenbach monitored concerns posted on social media about the level of disease during the H1N1 pandemic [10]. Scanfelt et al. analyzed the misunderstanding and misuse of antibiotics exposed through tweets [11]. Bosley et al. analyzed tweets pertaining to cardiac arrest and resuscitation [12]. In 2014, Preito et al. published an application of Twitter as a data set for healthcare quality surveillance. The authors data-mined more than 10 million tweets to predict health conditions in a population. This study was mainly focused on analyzing the incidence of flu, depression, pregnancy, and eating disorders in the Iberian Peninsula. The authors used geocoding information provided through the Twitter API to collect relevant tweets from the posted in Portugal and Spain, and then

determined a set of features in tweets that are likely written by users with the disease. They then created classes of tweets based on these features and tagged a subset of the tweets in order to create a training set. The classes consisted of positive, negative and undecided. The positive class consisted of tweets from Twitter users who are likely to have the disease. The negative class consisted of tweets that are users who are likely to not have the disease. And undecided is a class of tweets that could be both positive and negative. The results of Preto et al. study show usefulness in data-mining tweets to analyze health related topics. [14]

Our study will build on the literature that applies sentiment analysis to healthcare discussions. A study that provides relevant techniques for the analysis is *Using Twitter to examine smoking behavior and perception of emerging tobacco products.* In 2013, Zhu analyzed smoking behavior and the perception of emerging tobacco products through Twitter sentiment analysis. The author's objective was to create machine-learning classifiers in order to detect the sentiment towards tobacco products. The project collected 7,362 tobacco related tweets posted during the time from December, 2011, to July, 2012. Each tweet was classified as relevant or non-relevant. Additionally, the tweets were classified as either expressing positive or negative sentiment toward the use of a tobacco product. The authors found that there was a high positive sentiment[7] towards emerging tobacco products such as hookah and e-cigarettes. The authors also noticed disconnect between the use of products and their health effects. They concluded that the results demonstrate an

[7]The authors gave an example of a positive tweet related to tobacco to be: "Beer pongg / hookah round 2 with my goons waddduppppp. I love when my parents rnt home!"

opportunity for tobacco education. This study shows that machine-learning classification of Twitter data has a promising potential for social impact [13]. The conclusion of this paper contributes to my motivation to apply sentiment analysis to other health-related topics.

This research study will investigate Twitter data surrounding recommendation changes. The methodology of the study will avoid limitations of previous research that attempted a similar task. The study *Twitter response to the United States Preventive Services Task Force Recommendations against screening with prostate-specific antigen* by Prabhu et al., investigated the public and media response to the release of both the draft and finalized USPSTF recommendations against prostate-specific antigen testing. The authors data-mined tweets that contained the term "prostate cancer" within the 24 hours after the recommendation announcements. They found that increased twitter activity surrounded the announcements, with anti-screening tweets and articles increasing most. The authors were surprised that less than 10% of the tweets expressed opinion. They concluded that the study period was too short for sentiment analysis. The data may have also lacked tweets from international users due to day of the week and time zone differences. The authors suggested a study of longer duration with a baseline analysis before a recommendation is announced. This would require anticipating recommendation releases [15]. The methodology of my research study addresses this limitation. Our data was collected two months in advance of the draft recommendations and includes data during and after the release of the final recommendations.

Why Breast Cancer Screening?

Recently, breast screening has become a controversial public health issue due to the concern that mammography[8] leads to too many false-positives[9] and too much overdiagnosis[10] for the lack of impact it has on breast cancer mortality. In order to minimize these side effects of screening, a growing body of professionals recommends that mammography should be limited to populations at higher risk. Risk is evaluated through gender, age, individual health condition and history, and family history. Jennifer Frost, M.D., medical director for the AAFP Health of the Public and Science Division, explains that false-positives are most often seen with women under the age of 50 and in women who screen annually rather than biannually [21]. Her statement was in agreement with the USPSTF recommendations that the general population should begin biannual screening at age 50 rather than 40, to minimize harms such as false-positives.

In June 2015, Harding et al conducted an ecological study to determine if areas with more mammography had lower breast cancer mortality. The study reported that there was no correlation between increased mammography and subsequent breast cancer mortality. The authors concluded that the lack of correlation indicates that mammography is causing a significant amount of over-diagnosis [20]. What is troubling is that doctors have no way of knowing who is being over-diagnosed and

[8] Mammography is the most common imagining tool for breast cancer screening in the US and other developed countries.

[9] A result that indicates a patient has the disease when they are disease-free.

[10] Overdiagnosis is a diagnosis of a"disease" that will never cause symptoms or death during a patient's lifetime. Overdiagnosis is a side effect of screening for early forms of disease. (National Cancer Institute)

who is having a life-saving diagnosis; no current method exists to determine the course of a non-invasive tumor. Thus, despite over-diagnosis, many professionals continue to advocate for younger women to partake in routine mammograms. [19]

Due to the disagreement among professionals, breast cancer medical associations have conflicting recommendations regarding when and how often a woman should get screened. In the United States, the three influential associations are the United States Preventative Task Force, the American Cancer Society, and the National Cancer Institute. In 2009, the United States Preventative Services Task Force (USPSTF) updated its recommendation from annual screening starting at age 40 to recommending screening for ages 50-75 biannually. The USPSTF also recommended that those who are 40-50 should only be screened under certain circumstances and for women 75+ screening is not necessary. Since this 2009 statement, controversy spread not only nationally but world wide as well. The National Cancer Institute and American Cancer Society (ACS) disagreed with the USPSTF by continuing to recommend that women should be screened annually starting at age 40 to as long as they remain healthy. On the other hand, in 2014, the Swiss Medical Board recommended against introducing new mammography programs and to phase out existing programs [25]. The lack of evidence supporting the differing opinions created a large investment in research dedicated to evaluating the harms and benefits of screening the general population.

For the past six years, the varying recommendations stirred ongoing debate. In response to criticism, the USPSTF further investigated the harms and benefits of mammography, particularly in the younger age category. In April 2015, the USPSTF released a draft of their recommendations, which withheld the recommendations they made in 2009. The ACS has begun to recognize agreement with the USPSTF announcement; In October 2015 changed their recommendations to begin annual screening at age 45 and biannual screening after age 55 [24]. In January 2016, the USPSTF released the finalized recommendation changes [23]. Our research study will examine Twitter data pertaining to breast cancer screening before, during, and after the changes between April 16, 2015 and April 15, 2016.

The changes in recommendations cause great confusion with the public [21]. Policy-makers and physicians should understand the degree of confusion, and positive and negative sentiment about the recommendations in order to better consult patient. Previously, data that contained how the public perceived recommendation changes were found through surveys and through doctors' second hand. In the age of social media, this information can be found through data-mining and sentiment analysis. In 2012, Lyles et al. conducted the first study to analyze Twitter data pertaining to breast cancer screening. This study analyzed tweets over a five-week period. The authors' methodology involved manual coding of the tweet category and sentiment. The authors found that 25% of the Tweets were patients discussing their personal experiences with cancer screening. These messages provided information such as negative sentiment about procedures. The authors suggested value and further

research in using public sentiment for designing better health initiatives. This research paper will contribute to the work done by Lyles et al by extending the period of time studied as well as applying the analysis to an important time period in breast cancer policy. [22]

Chapter 2

Overview of Method

As discussed in Chapter 1: *Why Twitter*, access to the Twitter API is the foundation of this research study. In this section, I will discuss how we chose to mine tweets from the Twitter API, and then give an in depth explanation of the method used to analyze the Twitter data set. As shown in Figures 1 and 2, the methodology consists of two phases: Phase 1 is the process of building classifiers in WEKA that predict the relevance, sentiment, and category of a tweet. Phase 2 is the process of using the classifiers to predict the sentiment and categories of tweets over the course of a year.

Figure 1

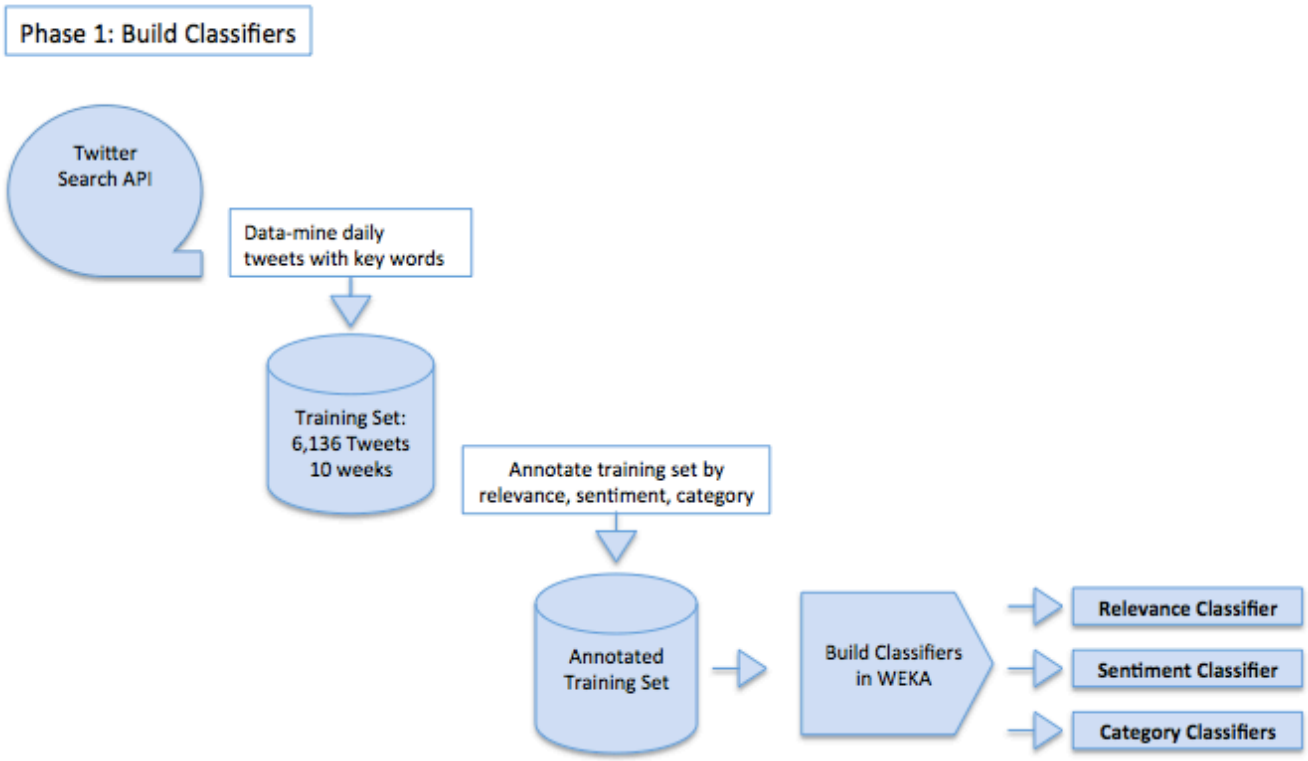
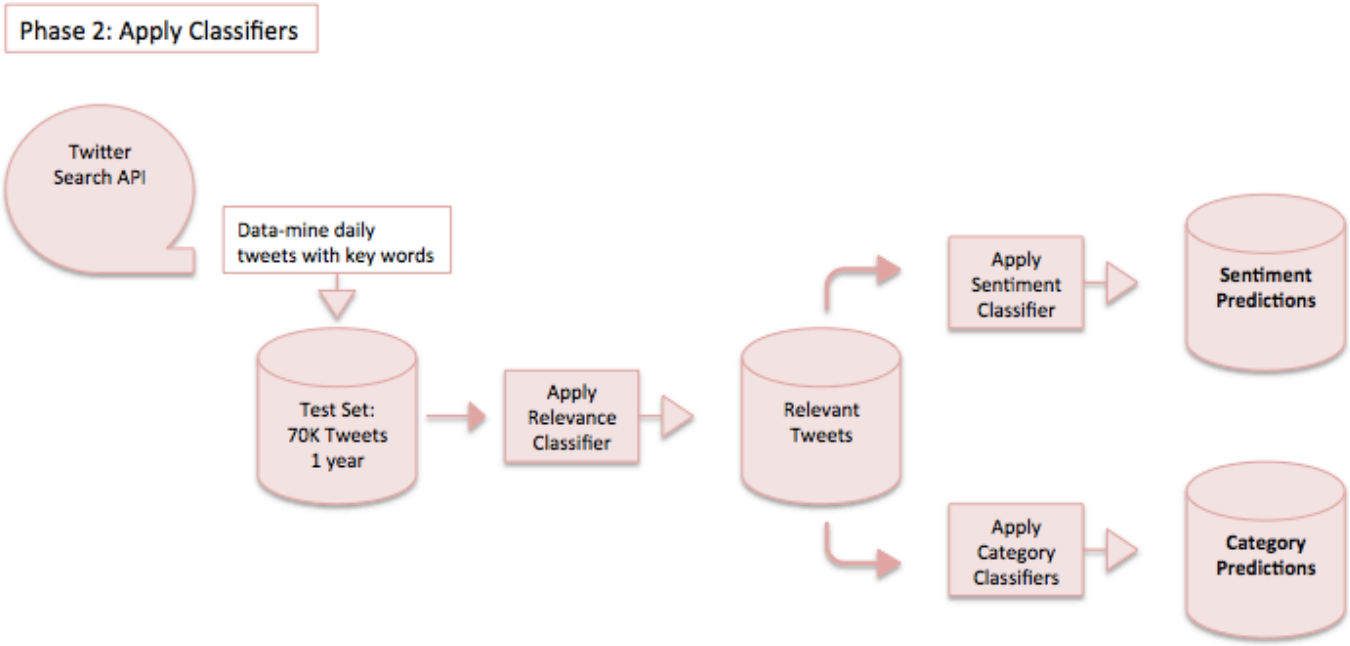


Figure 2



Phase 1: Data Collection, Annotation of Training Set, and Building Classifiers

Data Collection

Twitter provides three ways to mine tweets: the Twitter Search API, the Twitter Streaming API, and the Twitter Firehouse [30]. The Twitter Search API enables users to poll up to 5,000 of the most recent tweets from a particular user(s) or tweets with specific keyword(s). Twitter's Streaming API enables a user to retrieve a sample of tweets in real-time based on certain criteria such as keyword(s), location(s), or username(s). The Twitter Firehouse is a way of retrieving all tweets in real-time that can be accessed through GNIP or Datasift.

This study retrieved data through the Twitter Search API. Although the study would have benefited from a real-time engagement with Twitter data provided through the Twitter Firehouse or Streaming API, neither sources were accessible given our available resources; the Twitter Firehouse is expensive to access and the Streaming API requires a full-time listener to retrieve tweets. In order to still accomplish the task of mining a stream of tweets over a given period of time, we collected tweets by creating a cron[11] script. Everyday at 14:00, the cron job collected the most recent tweets and added the tweets to a local SQLite[12] database. To ensure that tweets are relevant and a single tweet is not retrieved multiple times, the script only added tweets that matched the following requirements: the tweet did not already exist in the database and the tweet contained a keyword from Table 1.

Table 1: Keywords[13]

"mammogra"
#mammogram
#mammography
#thermography
"breast thermogra"
"breast cancer screening"
"breast screening"
#breastscreening
#breast #screening
#3Dtomo
#tomosynthesis
#3dmammogram

[11] A cron is a time based job scheduler in Unix-like operating systems [32].
[12] A relational database management tool in C programming library [31].
[13] We compiled a list of keywords by conducting research of the most relevant terms, hash-tags, and words used in the online dialogue about breast cancer screening. The Twitter dialogue was our main focus, however current events concerning the use of new technologies were also used to gage the use of terms in future discussions on Twitter. (For example, tomosynthesis was being introduced in many locations at the start of this study)

"3D mammogram"
"3-D mammogram"
tomosynthesis

The data collection started in January 2015. We used the tweets that were collected from January 2015 through March 2015 to build a training set[14]. The training set consisted of 6,136 tweets. Starting on April 16, 2015, all tweets that were collected were added to a test set.[15] The test set consists of 77,387 tweets from April 16, 2015 to April 15, 2016.

Annotation of Training Set

Machine learning requires the use of a training set. A training set is a data set used to identify potentially valuable patterns in data in order to make predictions on the observed data set, otherwise known as a test set. In this research study, a training set was developed in order to create classifiers in WEKA. During the winter of 2015, we collected approximately 500-1,000 tweets per week for 10 weeks. Each week we read and categorized every tweet to a relevance, sentiment, and category. The annotation process is described in more detail below.

In the initial phases of the study, we investigated language and messages being used in the Twitter dialogue pertaining to breast cancer screening technologies. We concluded from this investigation that there are three interesting and potentially useful characteristics of the observed tweets: the relevance, the sentiment, and the

[14] The training set was created in order to conduct create machine-learning automated classifiers in Weka described in *Annotation of Training Set.*

[15] *The test set is the list of tweets that will be the focus of this study.*

category of the user experience. Table 2 shows the annotation options for each tweet. (Note that if a tweet were determined to be not relevant it would be annotated not relevant for sentiment and category as well.)

Table 2: Annotations

Relevance	**Sentiment**	**Category**
Not Relevant	Not Relevant	Not Relevant
Relevant	Positive Negative Neutral	Advertisement Awareness Controversy Personal Other

Relevance

Determining relevance of a tweet was necessary because the keyword data poll has the potential of collecting tweets that may contain a word in the keyword list, but is not relevant to the use of breast cancer screening technologies. For example, the following tweet contains the term "breast cancer screening" and is not relevant to the observed topic:

"the classic Mosque design in particular would make great breast cancer screening clinics."

Of the 6,136 tweets in the training set, 781 of the tweets were determined to be non-relevant.

Sentiment

As discussed in Chapter 1, this study builds off previous sentiment analysis research. Prior sentiment analysis studies suggest the useful technique of characterizing sentiment as positive, negative, or neutral. We agreed with the body of literature that supported this method. However, the rigidness of three types of sentiment can bring about difficulties with Twitter data. The first limitation we discovered is that a tweet can contain both negative and positive sentiment as exhibited in the following tweet:

My first ever #mammogram #mammograms #mammogramssavelives #scared

The Twitter user suggested positive sentiment towards mammography with #mammogramssavelives. Meanwhile, she expressed negative sentiment with #scared. Another example of a tweet with both positive and negative sentiment is:

Ladies make sure you keep your #Mammogram Appt. Mine might just have saved my life! About to start treatment for Early Stage BC next week :(

The tweet expressed positive sentiment towards the use of mammography. However, the Twitter user was negative about starting treatment. These cases demonstrated the need for criteria to ensure tweets that express multiple emotions were being annotated consistently. Our criteria are outlined in Table 3. Following the criteria in Table 3, both tweets discussed above were labeled as positive.

Table 3: Sentiment Criteria

Positive	**Negative**	**Neutral**
The tweet expresses positive emotion towards medical practices involving breast cancer screening technology in detecting early-stage cancer.	The tweet expresses negative emotion towards medical practices involving breast cancer screening technology in detecting early-stage cancer.	The tweet has no emotion towards the use of breast screening technology in detecting early-stage cancer.
Example: *Just got annual #mammogram. Boob party! #CheckItOffTheList. What are you checking off today???*	Example: *Do any of you have an opinion on this? Pretty interesting and extremely disappointing?!?! Breast screening a waste?*	Example: *Dont be alarmed if your #mammogram is abnormal. 80 percent - 90 percent of abnormal mammograms are not #cancer.*

In addition, sentiment annotation required an in-depth understanding of the language used to discuss breast cancer screening technologies. Terms such as "over-diagnosis" or "false-positives" were most likely indicating negative sentiment of the Twitter user. We know this due to the common criticism that mammography causes too much over-diagnosis and too many false-positives. On the other hand, "get boobies smashed" was often an indication of positive sentiment; the term is commonly used to encourage women to schedule a mammogram. These terms in a different context could express alternative sentiment. Thus, it was important to understand the language specific to breast cancer screening prior to the annotation process.

Determining sentiment of a tweet can also be tricky when a tweet expresses sarcasm. Sarcasm is common in Twitter language. Although the reader may understand that the tweet is sarcastic, a machine-learning classifier is likely to

wrongly detect the sentiment because sarcasm can disrupt patterns found in the training set. For example, the following tweet may contain sarcasm making it a difficult tweet to annotate:

I was contacted by the South East London Breast Screening service yesterday. I told them I have had a double mastectomy and have no breasts.

The Twitter user may be telling the truth. But it is also likely that her comment was sarcastic, expressing annoyance with the South East London Breast Screening service. Our research team discussed how we should address sarcasm and determined that we read the text without assuming the user has a sarcastic tone. The example tweet above was labeled as neutral.

Category

In the study, *Using Twitter for Breast Cancer Prevention: an analysis of breast cancer awareness month*, Thackeray et al investigated how Twitter was being used during October[16] 2012 [33]. Thackeray used words in the Twitter users' profile description to categorize a user as an organization, individual, or celebrity. The study performed a content analysis of the tweets and compared content among the three categories. The results show that Twitter content varies among different types of users. Thackeray et al. conclude that their research methods were limited by categorizing based off keywords in profile descriptions. Our research addressed the results and

[16] October is breast cancer awareness month.

limitations of Thackeray's study.

Taking Thackeray's results into consideration, we decided to categorize tweets in order to account for the variance in content among the different types of users. However, rather than using user description, we used the content of the tweet to classify categories. The categories were determined based off what we interpreted was the purpose of the tweet. The categories that were most prevalent in the training set were advertisement, awareness, controversy, and personal. Tweets that did not fit into any of the categories were labeled as other. Table 4 provides descriptions for each category. If a tweet could fall into more than one category we annotated based on the following ranking: personal, advertisement, controversy, and awareness.

Table 4: Category Criteria

Advertisement	**Awareness**	**Controversy**	**Personal**
The tweet is advertising a mammography clinic, doctor, or event	The tweet is promoting early detection	The tweet is contributing to the dialogue about the controversy over breast cancer screening recommendations	The tweet is about a Twitter user's personal experience or reaction
Example: *Do you need a #mammogram are uninsured & live in #Hardin County TX? We can help! Call Ava at 409-384-2099 for more info! #SETXNews #ETHAN*	Example: *Mammography is proven to save lives get up to date with your breast cancer screening #BCA #mammogram #breast cancer*	Example: *Most women should not get yearly mammograms experts confirm*	Example: *Im here getting my Girls checked out!!!! #mammogram*

Building Classifiers

Preprocessing in Weka

After the annotation of the training set was complete, we built automated machine-learning classifiers using WEKA. The first step to building classifiers in WEKA is preprocessing the data. The purpose of preprocessing is to create the cleanest version of the data to yield the most accurate classifier. The preprocess phase also serves as a way to format data for classifying algorithms to read.

WEKA enables a user to preprocess data in several different ways. In the preprocess tab, WEKA provides several options for stemmers[17], tokenizers[18], and the choice of using stopwords[19]. Tokenizing and stemming builds a dictionary of attributes[20] of which the classifier will be built on. Due to the many preprocess options, preprocess methods vary across previous research of Twitter in Weka.

We identified the best preprocess method for classifying each class (relevance,

[17] "A stemming algorithm is a process of linguistic normalization in which the variant forms of a word are reduced to a common form" (xapian.org). Stemming is done through a filter in WEKA called StringtoWordvector.

[18] A tokenizer is a tool used to break down a string of words into a set of terms, words, or symbols that will be the input for future processing. There are several ways to tokenize data. You will need to experiment with your given dataset to decide what tokenizer produces the best classifier.

[19] Stopwords are words that are used so often in a language that they do not add meaning to the sentence. Eliminating stopwords from your list of attributes may produce better classifiers but also may not. There are many different stopword lists available online for several different purposes. The one used during preprocessing is a generic stopwords list. *Example, "the," "an," "to," are all stopwords in the generic list.*

[20] If you were to think of an instance as a statement or sentence, the attribute would be a word or clause that adds meaning to the instance. During the preprocessing phase, WEKA will tokenize (splice) each instance into a list of attributes. There are many different tokenizing methods.

sentiment, category) by observing the method that yielded the best results. Accuracy of results was determined by the percent correctly classified after a 10 fold cross-validation in the WEKA classifier tab. Table 5 outlines the preprocess method for each class.

Table 5: Preprocess per Class

Class	**Stemmer**	**Tokenizer**	**Stopwords**
Relevance	Snowball Stemmer	Word Tokenizer	Removed stopwords
Sentiment	Snowball Stemmer	Alphabet Tokenizer	Did not remove stopwords
Category	Snowball Stemmer	N-gram Tokenizer	Did not remove stopwords

Creating Prediction Model in Weka

After the preprocessing was complete, we used the classifier tab in Weka to test how different classifiers performed on the data. The performance was measured based on the percentage of tweets that were correctly classified in a 10-fold cross-validation. The classifiers we tested were J48, SMO, and NaiveBayes because they were most commonly used in similar studies in the literature. We found that SMO had the best results for all classes. Our goal for percent classified correctly was initially 75%. Figures 3-8 show that the performances of our classifiers exceeded our expectations.

Figure 3: Relevance SMO Classifier Performance

Correctly Classified Instances	5931	96.6591 %
Incorrectly Classified Instances	205	3.3409 %
Kappa statistic	0.8414	
Mean absolute error	0.0334	
Root mean squared error	0.1828	
Relative absolute error	15.0315 %	
Root relative squared error	54.8421 %	
Coverage of cases (0.95 level)	96.6591 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	6136	

Figure 4: Sentiment SMO Classifier Performance

Correctly Classified Instances	4711	88.4695 %
Incorrectly Classified Instances	614	11.5305 %
Kappa statistic	0.7945	
Mean absolute error	0.2541	
Root mean squared error	0.3245	
Relative absolute error	67.2535 %	
Root relative squared error	74.6589 %	
Coverage of cases (0.95 level)	97.3521 %	
Mean rel. region size (0.95 level)	66.9671 %	
Total Number of Instances	5325	
Ignored Class Unknown Instances	5963	

Figure 5: Advertisement SMO Classifier Performance

Correctly Classified Instances	5066	95.1362 %
Incorrectly Classified Instances	259	4.8638 %
Kappa statistic	0.7191	
Mean absolute error	0.0486	
Root mean squared error	0.2205	
Relative absolute error	26.4088 %	
Root relative squared error	72.7017 %	
Coverage of cases (0.95 level)	95.1362 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	5325	

Figure 6: Awareness SMO Classifier Performance

Correctly Classified Instances	4818	90.4789 %
Incorrectly Classified Instances	507	9.5211 %
Kappa statistic	0.7286	
Mean absolute error	0.0952	
Root mean squared error	0.3086	
Relative absolute error	25.8482 %	
Root relative squared error	71.9056 %	
Coverage of cases (0.95 level)	90.4789 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	5325	

Figure 7: Controversy SMO Classifier Performance

Correctly Classified Instances	4954	93.0329 %
Incorrectly Classified Instances	371	6.9671 %
Kappa statistic	0.7546	
Mean absolute error	0.0697	
Root mean squared error	0.264	
Relative absolute error	23.4707 %	
Root relative squared error	68.5235 %	
Coverage of cases (0.95 level)	93.0329 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	5325	

Figure 8: Personal SMO Classifier Performance

Correctly Classified Instances	5165	96.9953 %
Incorrectly Classified Instances	160	3.0047 %
Kappa statistic	0.774	
Mean absolute error	0.03	
Root mean squared error	0.1733	
Relative absolute error	20.6128 %	
Root relative squared error	64.2397 %	
Coverage of cases (0.95 level)	96.9953 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	5325	

Our classifiers out-perform similar studies found in the literature. As mentioned in Chapter 1, Zhu et al. conducted a sentiment analysis of tobacco products using Weka. Their most accurate classifiers had an accuracy of 84% [13] while our sentiment classifier had an accuracy of 85%. In Umadavi's study *Sentiment Analysis Using WEKA*, Umadavi compared the accuracy of two machine-learning algorithms in Weka, Support Vector Machine and Decision Tree. The most accurate models in Umadavi's study ranged from accuracies of 87% to 92% [4].

Our sentiment classifier likely has a lower accuracy compared to the other classifiers due to limitations discussed previously of sentiment analysis of Twitter. In addition, the sentiment classifier made predictions between three buckets[21],

[21] Bucket is a term used to describe the option a predictive model has for a prediction. The sentiment classification had three buckets: *positive*, *negative*, and *neutral*. The other predictive models had two

while the other classifiers only chose between two buckets, which increased the model's susceptibility to error.

Phase 2: Applying Classifiers to Data Set

From April 16, 2015 to April 15, 2016, 77,387 tweets[22] were mined from the Twitter Search API. Using WEKA, we applied the models built from the training set to make predictions on the test set. As shown in Figure 2, we first applied the relevance classifier to the test set. Then, we eliminated tweets that were predicted to be not relevant. The results of our relevant selection process produced a new test set of 69,704 tweets. We used the sentiment and category models on this new test set, in order to achieve our results.

Results

The results of the classifier predictions can be seen in Figures 9 and 10. The timeline in Figures 11 and 12 shows the dynamic results[23]. Figure 11 displays the results of sentiment over the course of the year from April 16, 2015 to April 15, 2016. Figure 12 displays the results of category. The focus of our analysis will be on how time and events effect changes in sentiment and category shown in Figures 11 and 12.

---

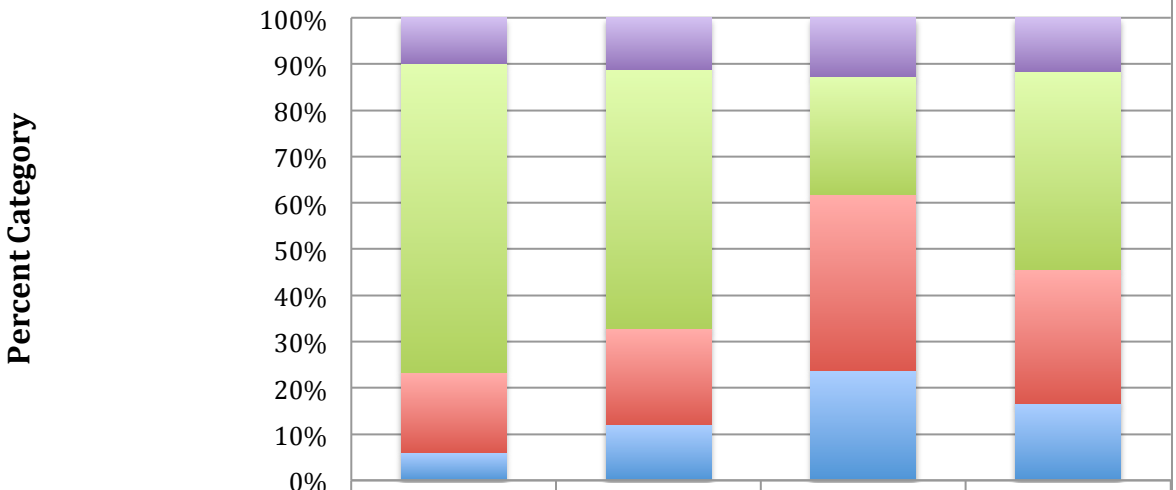
buckets to choose from. For example, the buckets for advertisement were *advertisement* and *not an advertisement*.

[22] This was the test set.

[23] The images were created through an open sourced java script and are available for an interactive view at http://passumpsic.dartmouth.edu/skikut/index.html. The toggle in the upper right corner of the page guides the user to sentiment, category, and baseline graphs. The graphs can be enlarged using the gray timeline at the top of the page. Click and drag an area of the timeline to zoom into particular time frames. To see the individual data point, hover cursor over point of interest and a hover box should appear with category/sentiment breakdown as well as top five words of that day.

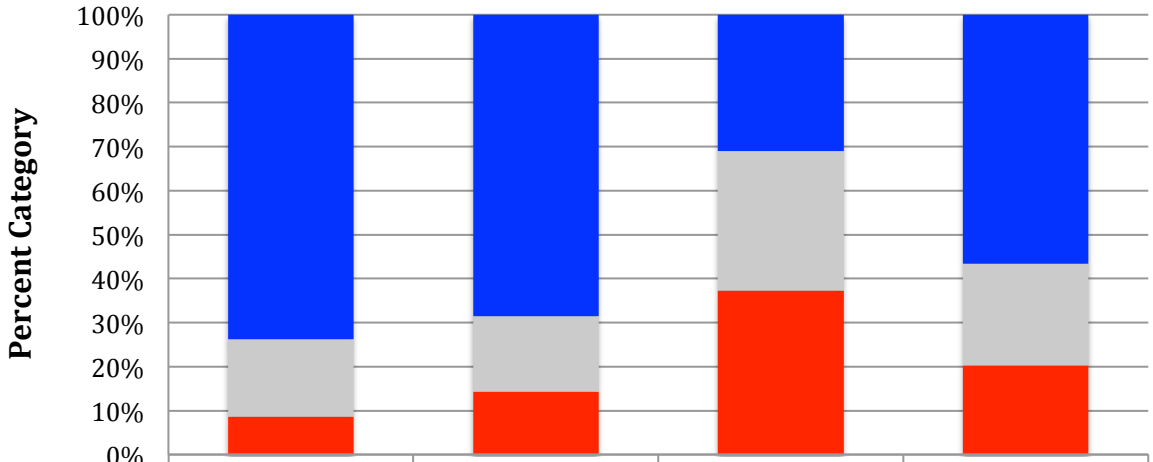
Figure 13 shows the results of a data pull unrelated to breast cancer screening, serving as a control for the study.

Figure 9: Comparison of Sentiments by Percent Category



	Negative	Neutral	Positive	Grand Total
Sum of Personal	789	903	2201	3893
Sum of Controversy	5331	4520	4430	14281
Sum of Awareness	1376	1673	6623	9672
Sum of Advertisement	483	974	4107	5564

Figure 10: Comparison of Categories by Percent Sentiment



	Sum of Advertisement	Sum of Awareness	Sum of Controversy	Sum of Personal
Positive	4107	6623	4430	2201
Neutral	974	1673	4520	903
Negative	483	1376	5331	789

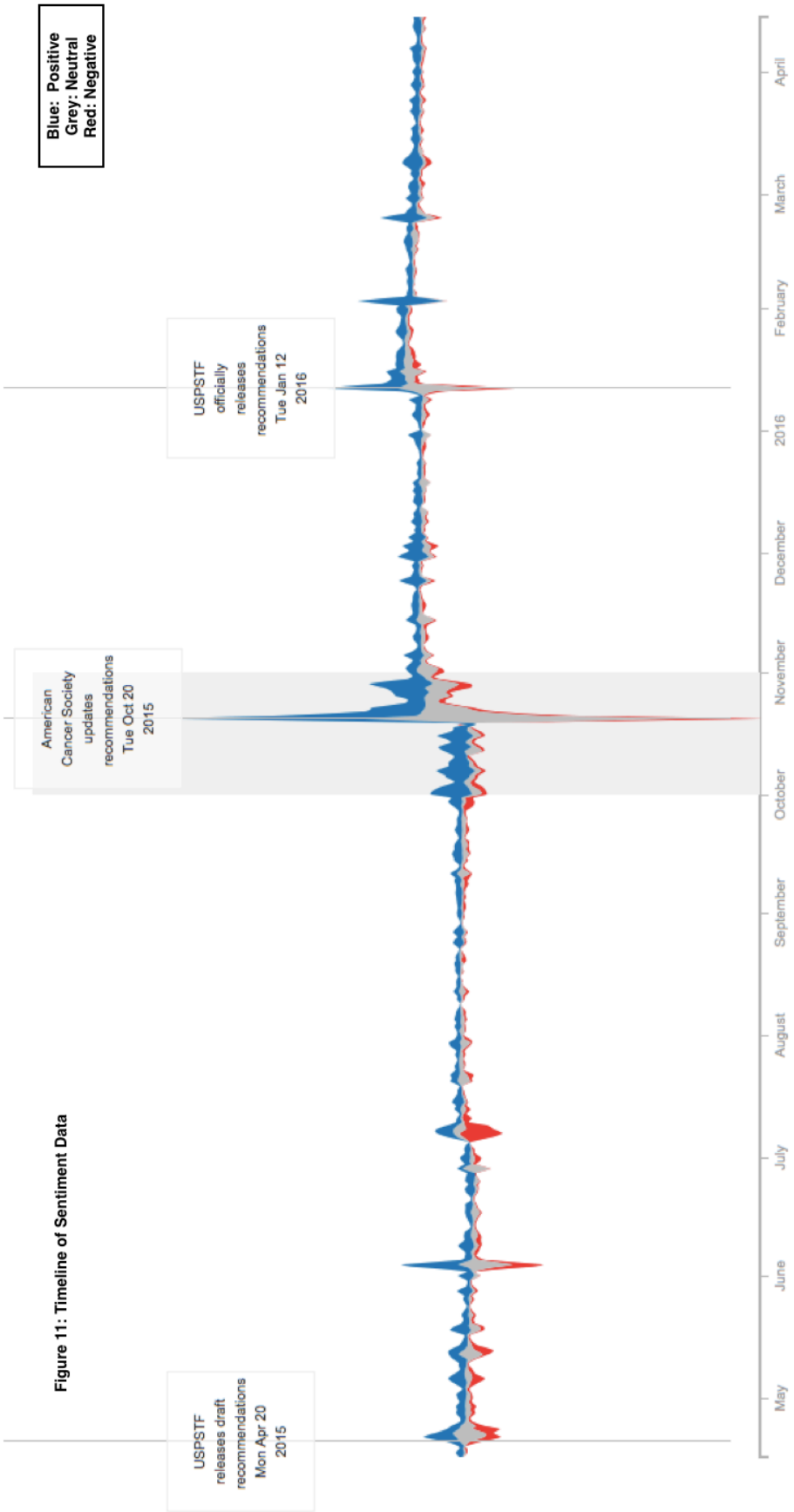


Figure 11: Timeline of Sentiment Data

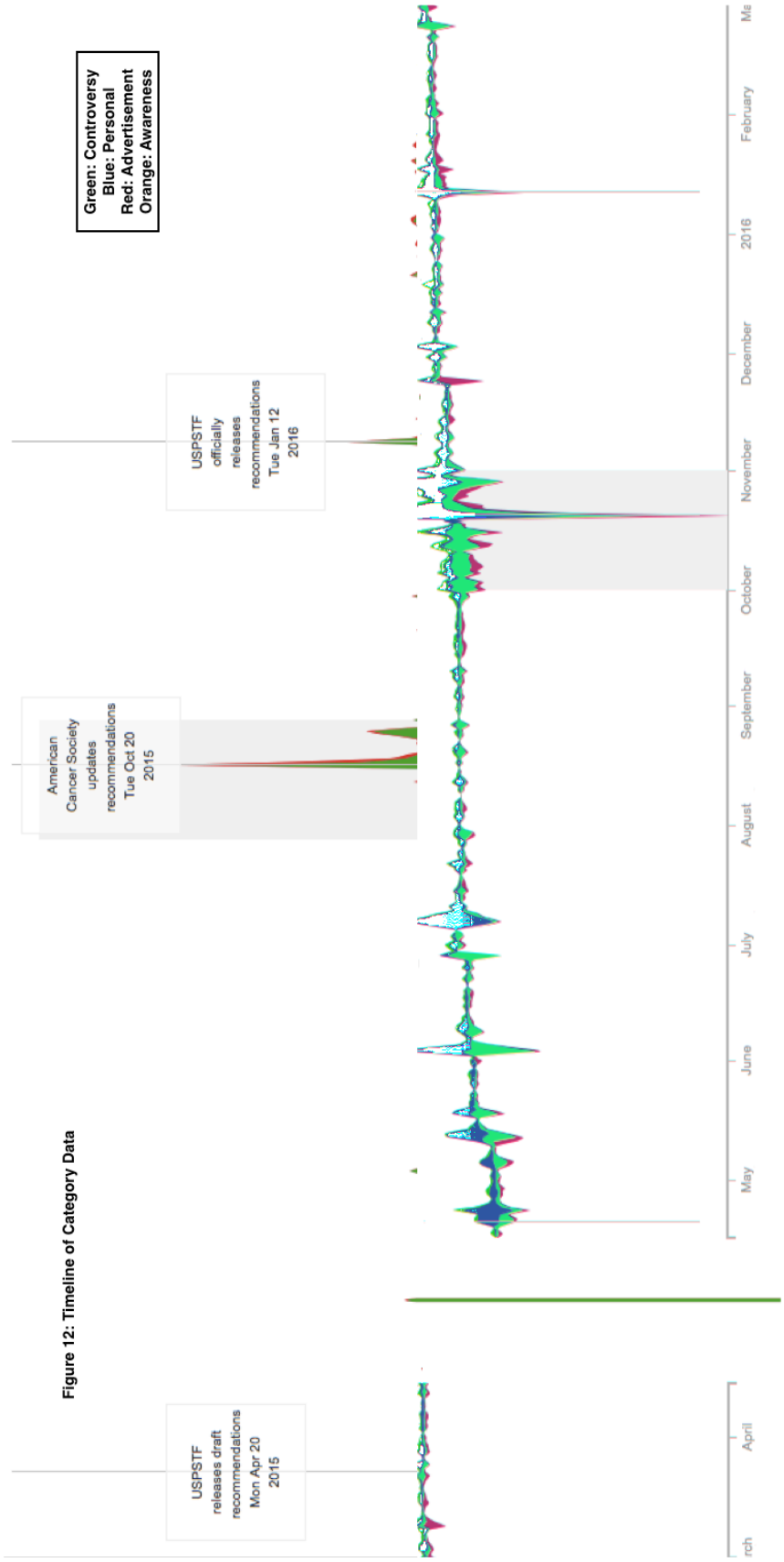


Figure 12: Timeline of Category Data

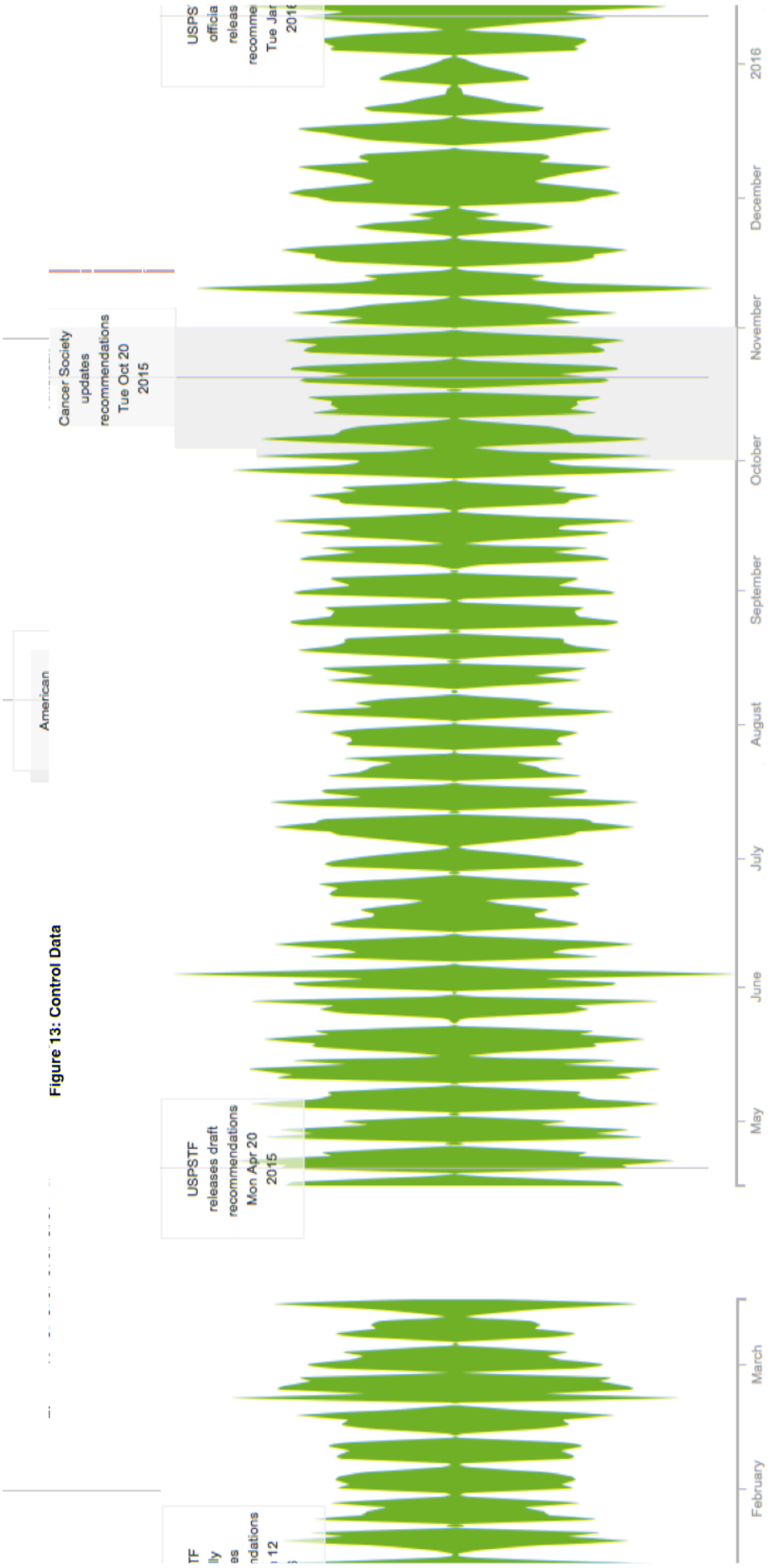


Figure 13: Control Data

The distribution shown in Figures 9 and 10 (same data shown in two different ways) had a chi-square statistic of 4882.2055. The result is statistically significant with p-value < 0.00001.

Figures 9 and 10 illustrate that the controversial discussion concerning breast-screening recommendations contributed most to the negative and neutral discussions on Twitter. Awareness, Advertisements, and Personal messages all have the greatest amount of positive content, which is not surprising. Neutral messages range from 15% to 30% of content per category. This difference is small compared to positive content, which ranges from 30% to 75% of content per category. And negative content ranges from 9% to 36% of content per category. We can conclude from this that the category of the tweet affects the likelihood that it is a positive or negative tweet.

Figure 9 shows that controversy has the greatest number of tweets. We hypothesize that this is due to the announcement and responses to recommendation changes. In addition, we hypothesize that the high volume of awareness content is due to the increased use of Twitter during breast cancer awareness month. These hypotheses are confirmed in Chapter 3.

Figure 13, the control data, suggests that the day of the week may affect Twitter activity: people are more likely to tweet about observed topics on weekdays. Figures 11 and 12 support this hypothesis. Thus, in Chapter 3, we will conduct a

quantitative analysis of how day of the week may affect the volume in order to control for day of the week in the analysis. In addition, Figures 11 and 12 show that volume, sentiment, and category, may be significantly impacted by recommendation changes. We will investigate the impact of events on Twitter data in Chapter 3.

Chapter 3

Overview of Method

In this chapter, we determine the effects of events on the breast cancer screening Twitter dialogue. First, we examine if the day of the week impacts the volume of Twitter data per day and the likelihood of a tweet being positive, negative, or neutral per day. Second, we analyze how Twitter data is affected by breast cancer awareness month (BCAM). Third, we evaluate if the American Cancer Society (ACS) recommendation change corresponds with significant changes in Twitter sentiment and content. Last, we will test whether changes in Twitter data surrounding the ACS recommendation change occur in the Twitter data surrounding the USPSTF finalized recommendation. We conclude by suggesting that the similarity of the Twitter data changes surrounding these two events demonstrates that specific Twitter trends occur surrounding recommendation changes.

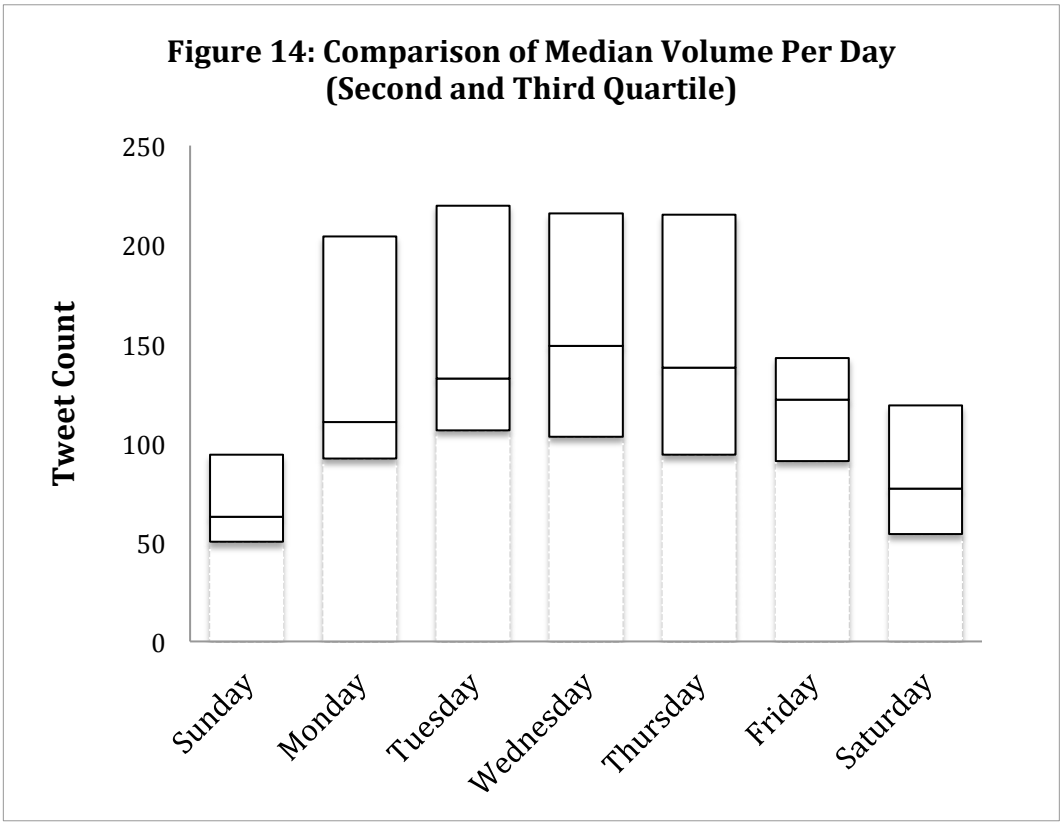
Effect of Day of Week on Twitter Data

After conducting a visual analysis of the Twitter data timeline shown in Figures 11 and 12, we hypothesized that the inconsistency in volume per day (when no major events occur) is due to the effect of day of the week: Twitter users may be more active on certain days of the week. This information is necessary to consider for future analysis in order to accurately assess how Twitter mood and content responds to a specific event. We conducted a quantitative analysis to investigate the effect of day of the week on volume and sentiment.

We calculated the median volume per day shown in Figure 14 and Table 7. For the median, we used a 25-75% range. The results indicate that volume differs by day of the week. Twitter activity is significantly greatest midweek, Tuesday through Thursday. Thus, when analyzing tweet count surrounding a particular event we will control for the effect of day of the week by using the daily medians as the baseline.

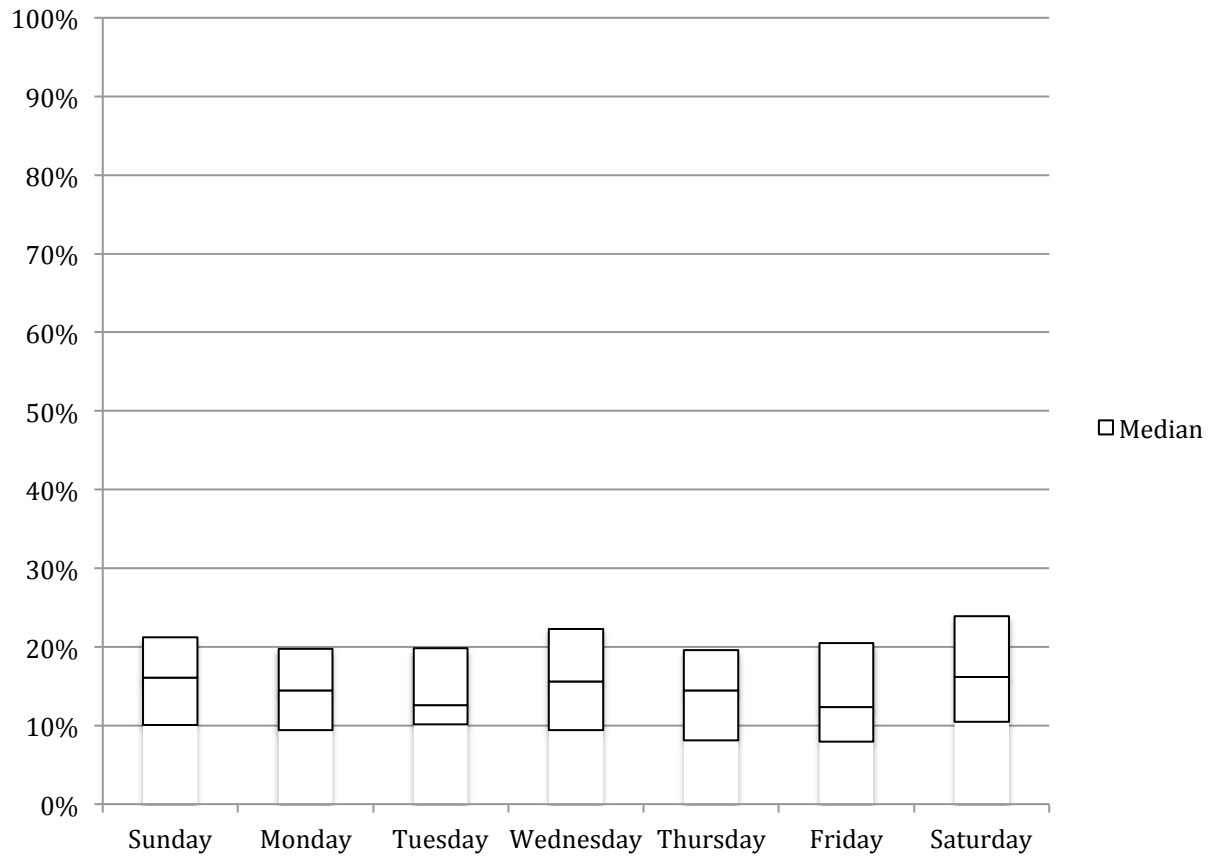
Table 7: Median Tweet Count Per Day

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Quartile 1	50	92	107	103	94	91	54
Median	63	110.5	132.5	149	138	122	77
Quartile 3	94	204	220	216	215	143	119



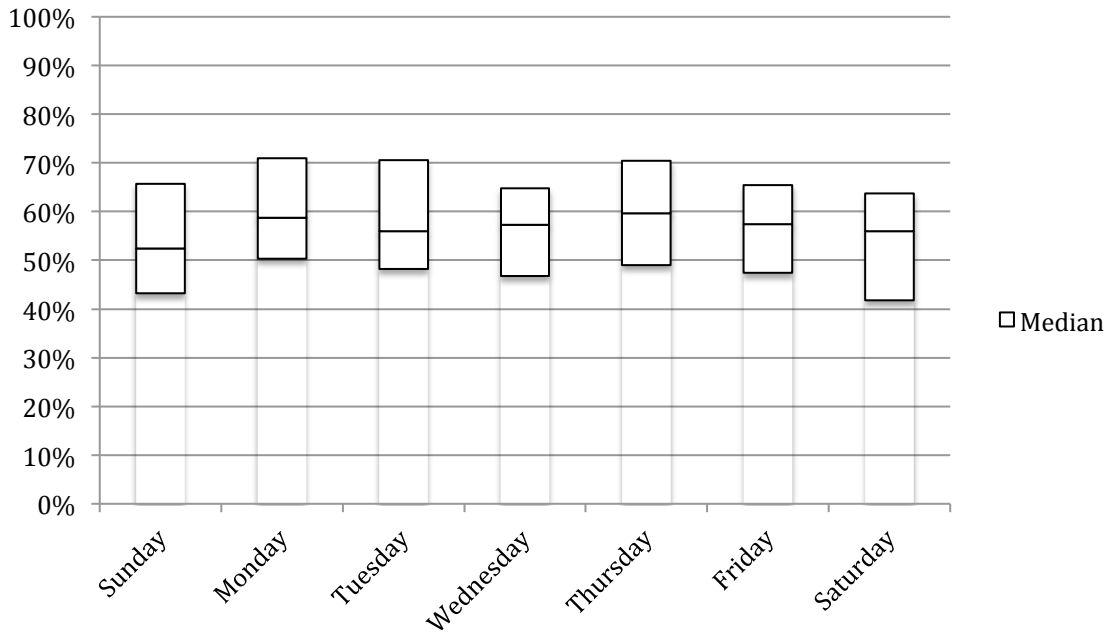
Next, we calculated the median, with a 25-75% range, percent positive and percent negative per day to identify if Twitter mood is affected by the day of the week.[24] Our results shown in Figures 15-16 indicate that sentiment is not significantly impacted by the day of the week. Day of the week will not be controlled for in further sentiment analysis.

Figure 15: Median Percent Negative Per Day (Second and Third Quartile)



[24] Results for neutral percent per day are not shown. We determined that neutral sentiment is not a characteristic of mood change during different days of the week.

Figure 16: Median Percent Positive Per Day (Second and Third Quartile)



Event Analysis

Breast Cancer Awareness Month

In 1985, the American Cancer Society (ACS) partnered with Imperial Chemical Industries to create what is now known as Breast Cancer Awareness Month (BCAM), taking place during October. The goal of BCAM was to promote mammography as a screening device to decrease breast cancer.[25] As shown in Figures 11 and 12 (Chapter 2), Breast Cancer Awareness Month (BCAM) creates a significant amount of activity on Twitter. We conducted an analysis to quantify the effect BCAM has on the Twitter dialogue pertaining to breast cancer screening. First, we quantified the

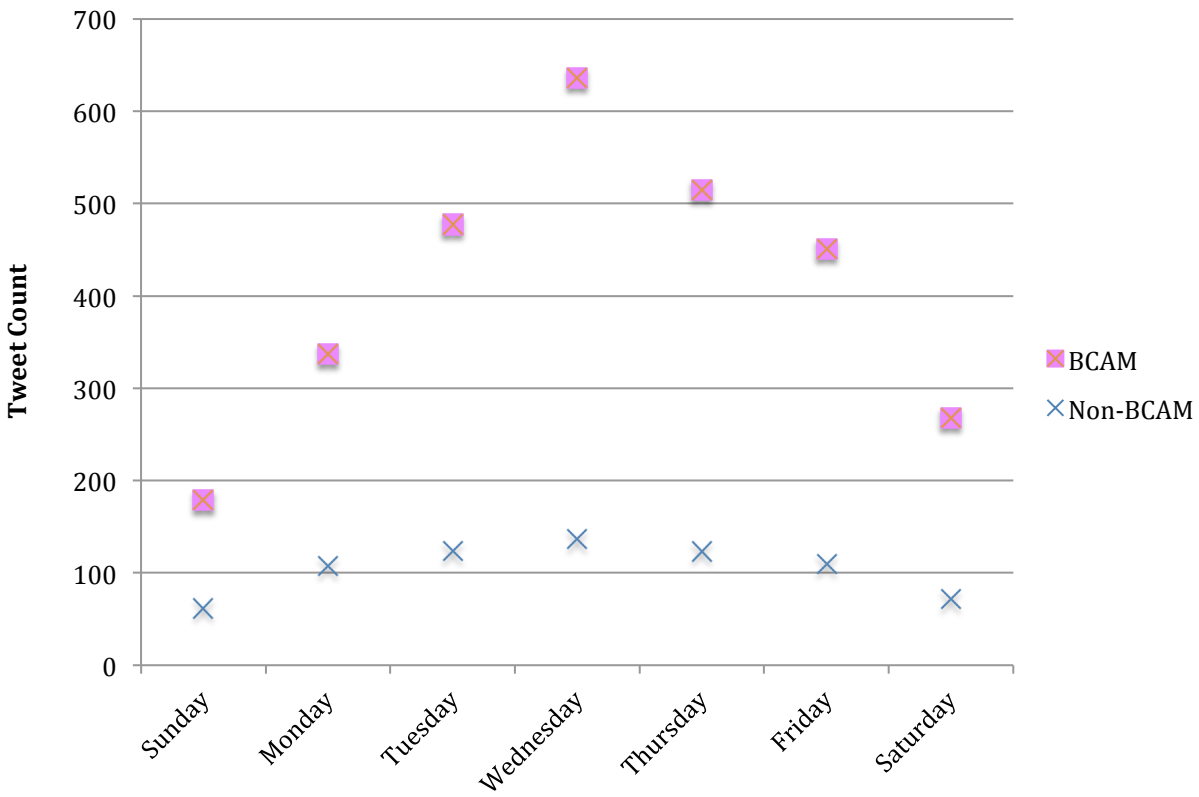
[25] https://web.archive.org/web/20110716123431/http://www.nbcam.org/about_faq.cfm

effect of BCAM on the count of tweets per day. Next, we examined how the increase of count per day affects the count of tweets in each category. And finally, we calculated the effect of BCAM on Twitter sentiment.

BCAM Effect on Volume

We calculated the median volume per day during BCAM and compared the results to the median of non-BCAM data. The results are shown in Figure 17 and Table 8. The non-BCAM medians range from 62 tweets on Sundays to 137 tweets on Wednesdays (Table 8). The median drastically increased during BCAM to range from 179 tweets on Sundays to 636 tweets on Wednesdays. The results show that Wednesdays are

Figure 17: BCAM Median Count



the most active day for Twitter users regardless of the effect of BCAM and BCAM has a significant impact on the amount breast cancer screening is discussed on Twitter.

Table 8: Median Count Per Day during BCAM

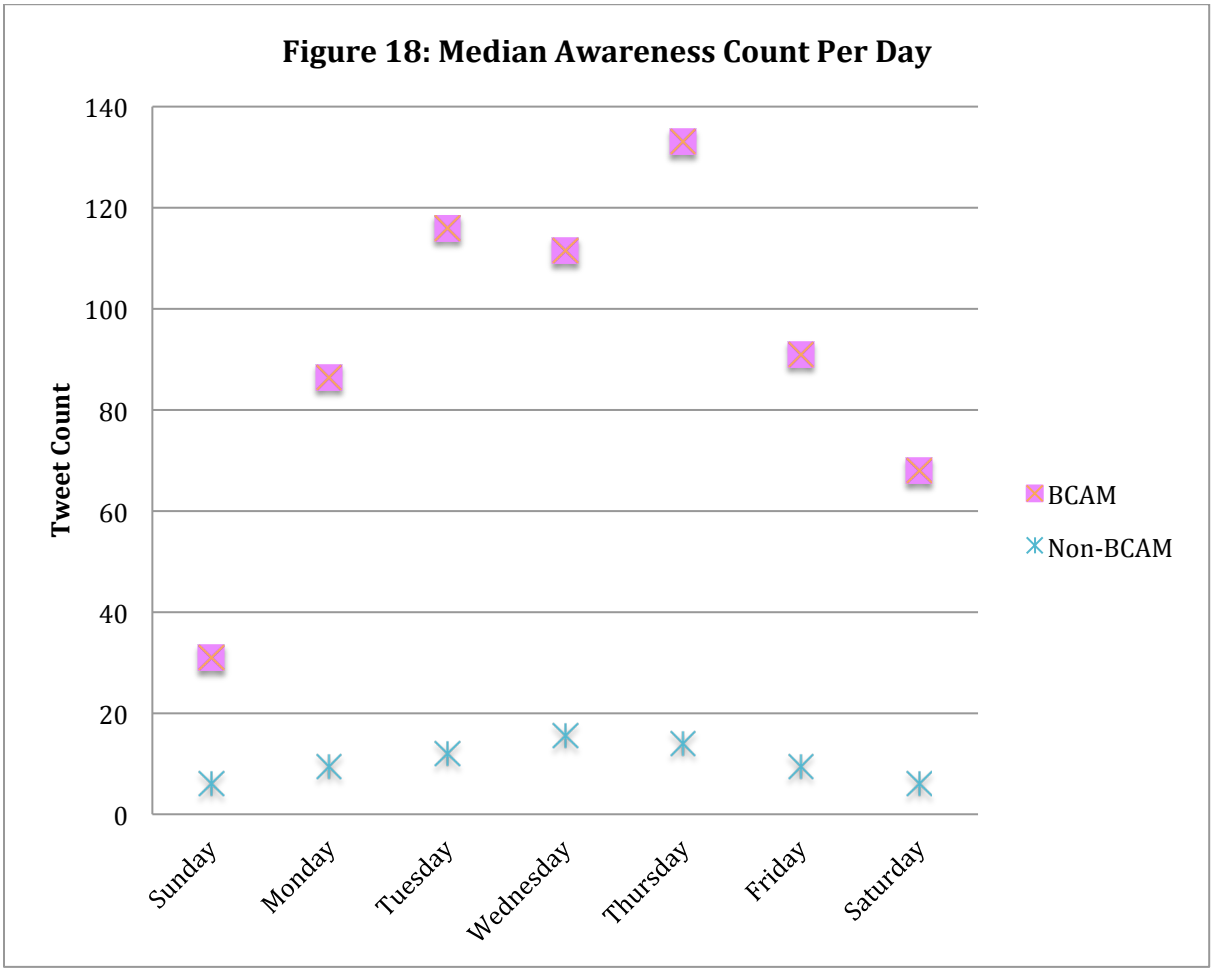
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Non-BCAM Median	62	107	123.5	137	123	109.5	72
BCAM Median	179	337	477	636	515	451	268

BCAM Effect on Category

We determined the effect BCAM had on the count of tweets per day in each category. Our findings, displayed in Figures 18-21, show that Twitter activity increased in all categories. Additionally, most Twitter activity took place during the week. The categories varied among which day activity was greatest. These findings may suggest tendencies of Twitter users and behavior that are specific to BCAM.

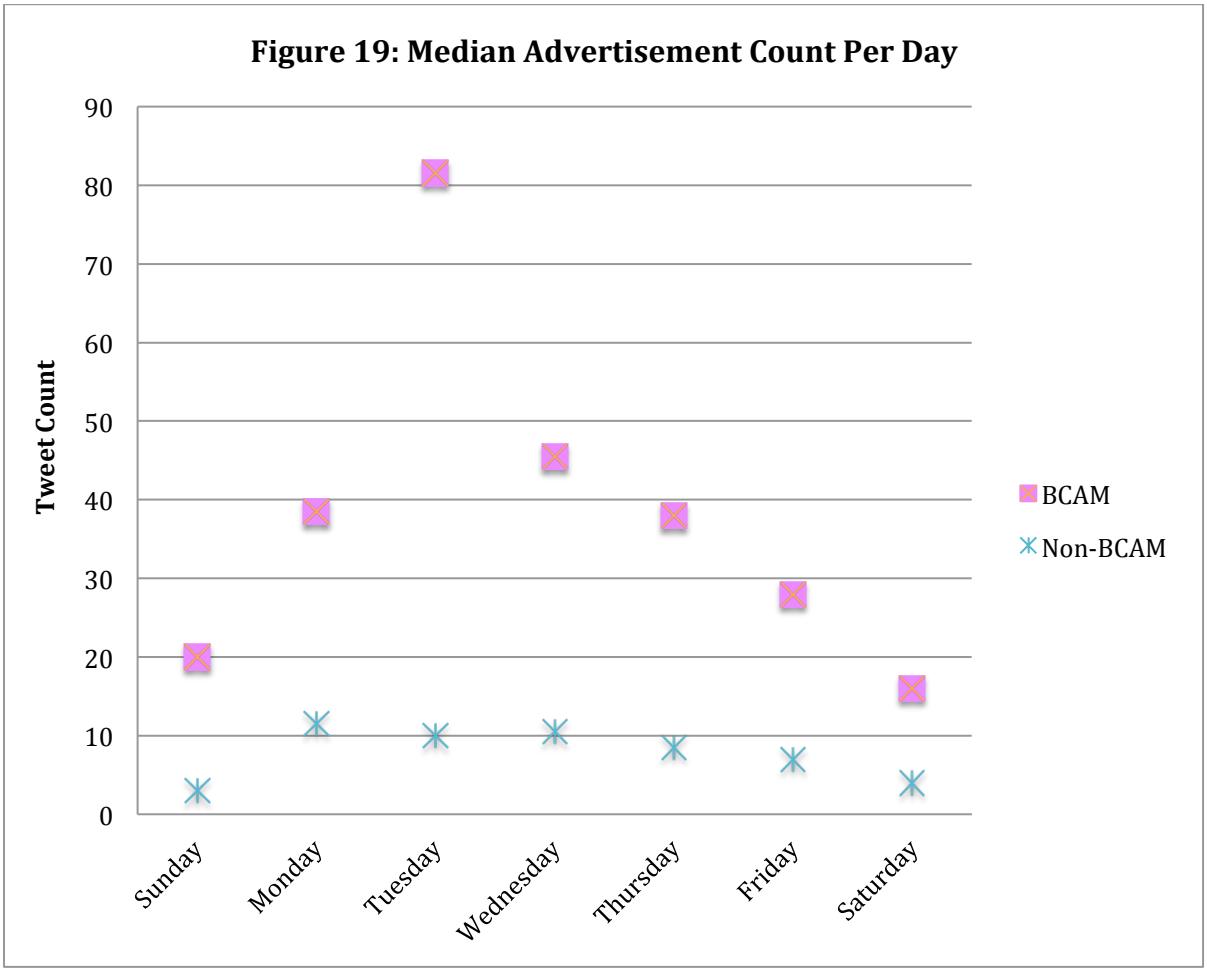
As expected, awareness activity had the greatest increase of tweets during BCAM (Figure 18). The non-BCAM median of count per day ranges from 6 to 16 tweets per day. During BCAM, the median count per day ranged from 31 to 133 tweets per day. Thursday had the greatest increase with a median of 133 tweets. This finding is not surprising given that awareness activity is greatest on Thursday in the baseline as well.

Figure 18: Median Awareness Count Per Day

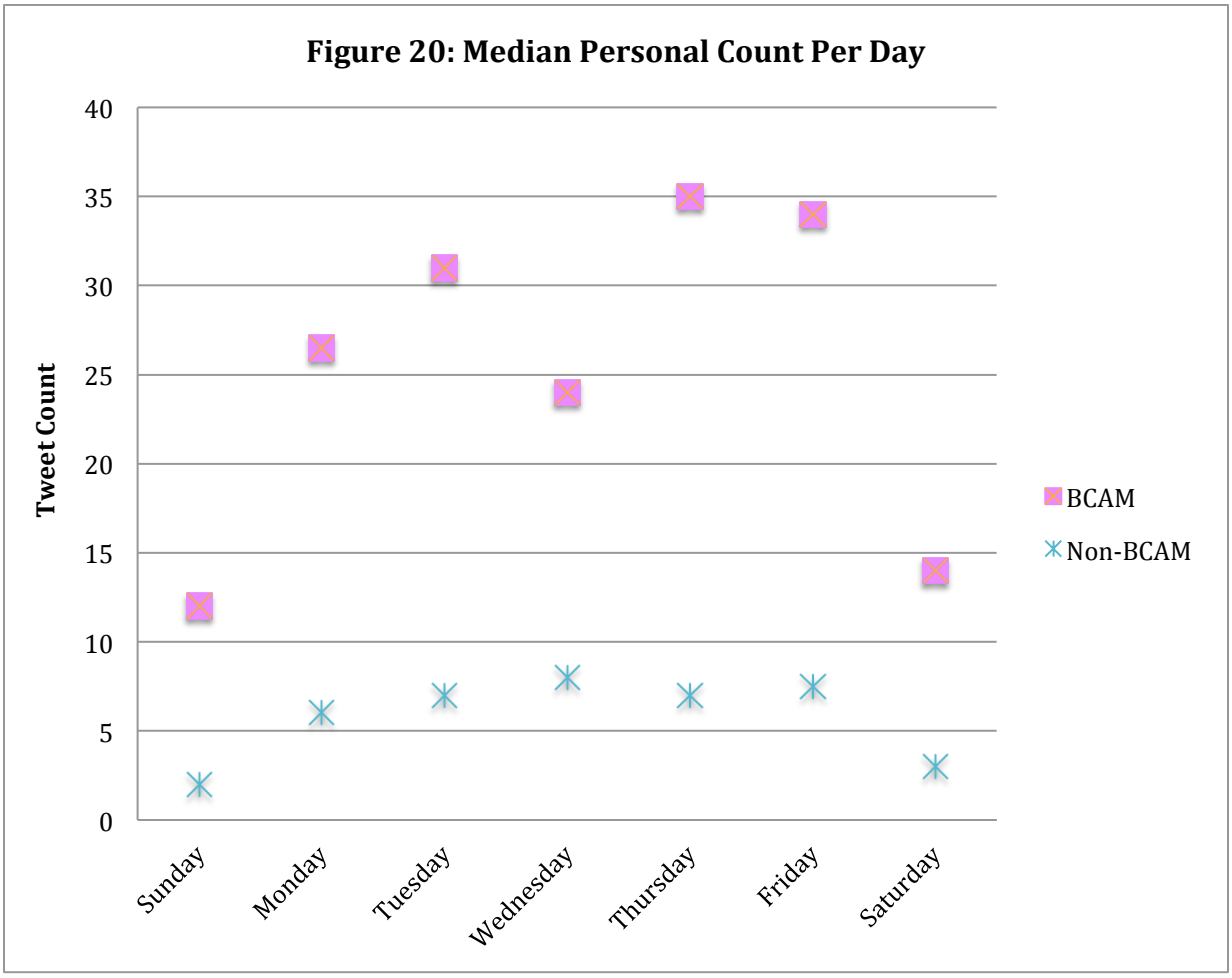


Advertisements had a three-fold increase in tweet count per day during BCAM (Figure 19). The non-BCAM median count per day ranged from 3 to 12 tweets per day. The BCAM median count per day ranged from 16 to 82 tweets per day. Unlike non-BCAM months, the most active day for advertisements during BCAM was on Tuesday. Future work could investigate if the increased advertisement activity on Tuesday correlates with clinical trends.

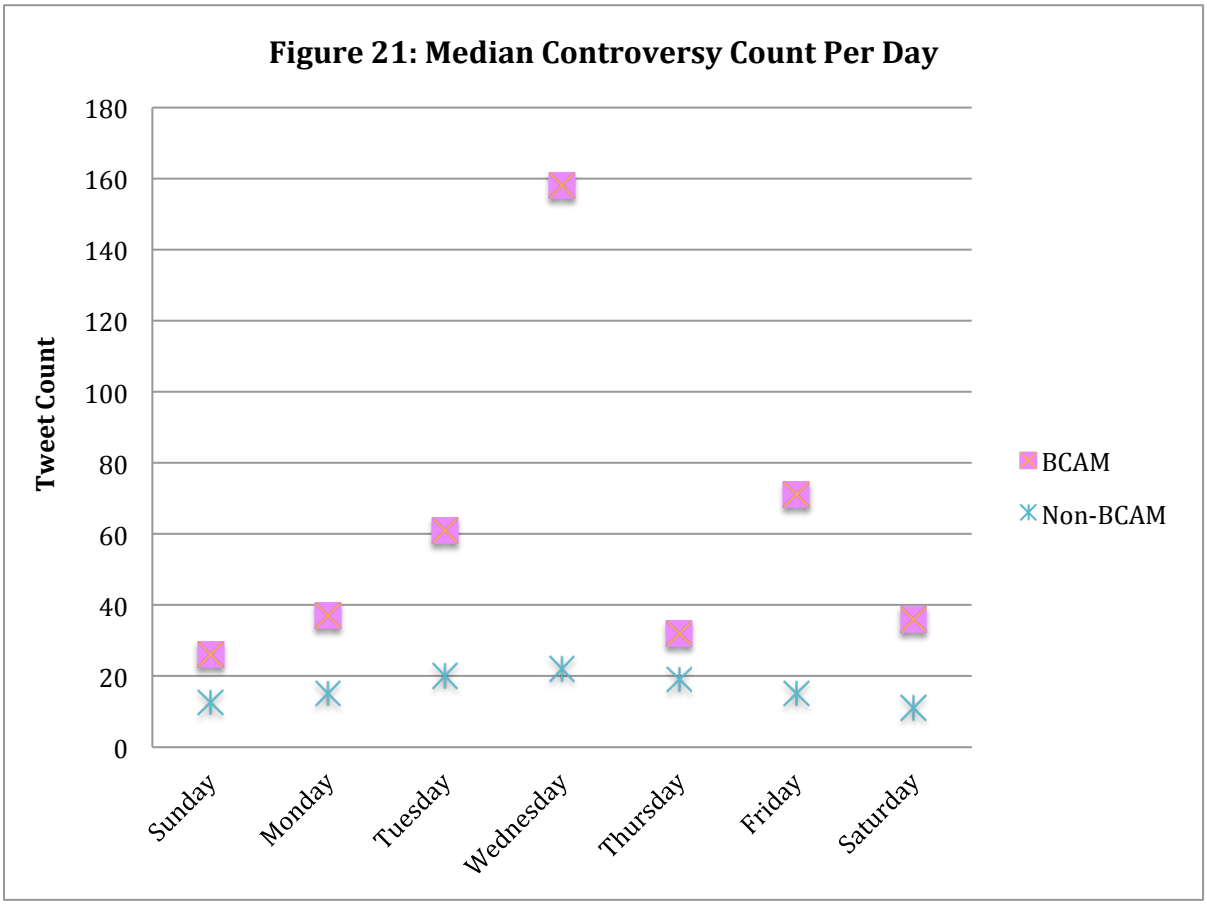
Figure 19: Median Advertisement Count Per Day



Personal tweets increases during BCAM as well (Figure 20); the median count per day ranges from 12 to 35 tweets per day. During the other months of the year, the median count per day ranges from 2 to 8 tweets per day. Thursday was the most active day for personal tweets in BCAM.



Tweets related to breast cancer screening controversy increased during BCAM (Figure 21); the median count per day ranges from 26 to 158. During the non-BCAM months, the median count per day ranges from 11 to 22 tweets per day. Wednesday had the greatest increase in controversy activity during BCAM, with a median that is 136 tweets above the baseline median.



BCAM Effect on Sentiment

After determining the effect of BCAM on the count of tweets per day, we evaluated how and if sentiment changed during BCAM. We hypothesized that BCAM would increase positivity and decrease negativity: BCAM serves to increase awareness and

participation in breast cancer management and health. Our findings confirmed that Twitter activity responds to BCAM in a more positive and less negative mood: the median increase from the baseline for percent positive during BCAM is 6% and the median decrease from the baseline for percent negative during BCAM is 4%.

In conclusion, our results indicate that tweeting about breast cancer screening is more likely to occur during BCAM. In addition, the dialogue during BCAM is more positive and less negative. These results raise the question of whether an event that occurred during BCAM disrupted BCAM sentiment trends. In the next section, we answer this question by conducting a sentiment analysis of Twitter data surrounding the ACS recommendation change.

American Cancer Society Recommendation Change

As discussed in Chapter 1, since the USPSTF updated its recommendation in 2009, many studies, policy changes, and events have continued to call question to how screening programs should be implemented. This research study uses Twitter to analyze how information about breast cancer screening is being communicated on Twitter and the effect that the information has on the relevant Twitter dialogue. We chose to conduct a detailed event analysis of Twitter data surrounding the ACS recommendation change because the event is arguably the most controversial story of the past year. Unlike the USPSTF that has been encouraging cut backs to screening programs for several years, the ACS has just recently acknowledged a change in position. In addition, the ACS recommendation change was released during BCAM. As we have shown, Twitter activity increases in volume and positivity

during BCAM. Therefore, the timing of the ACS recommendation change is important to our analysis. We hypothesize that the recommendation change decreases positive sentiment due to the increased content about the announcement. We also expect a lag time for sentiment to return to the expected percent after the recommendation. In this section, Twitter sentiment as well as the content of tweets is analyzed surrounding the ACS recommendation change. We will use the results to test whether the USPSTF finalized recommendation had the same effect on Twitter data. If so, we conclude that specific Twitter trends occur surrounding recommendation changes.

ACS Recommendation Effect on Sentiment

Figures 22-24 display the positive, negative, and neutral sentiment two weeks before and after the ACS recommendation. We identify the change in sentiment by calculating the mean sentiment of the data two weeks before the event and two weeks after the event.

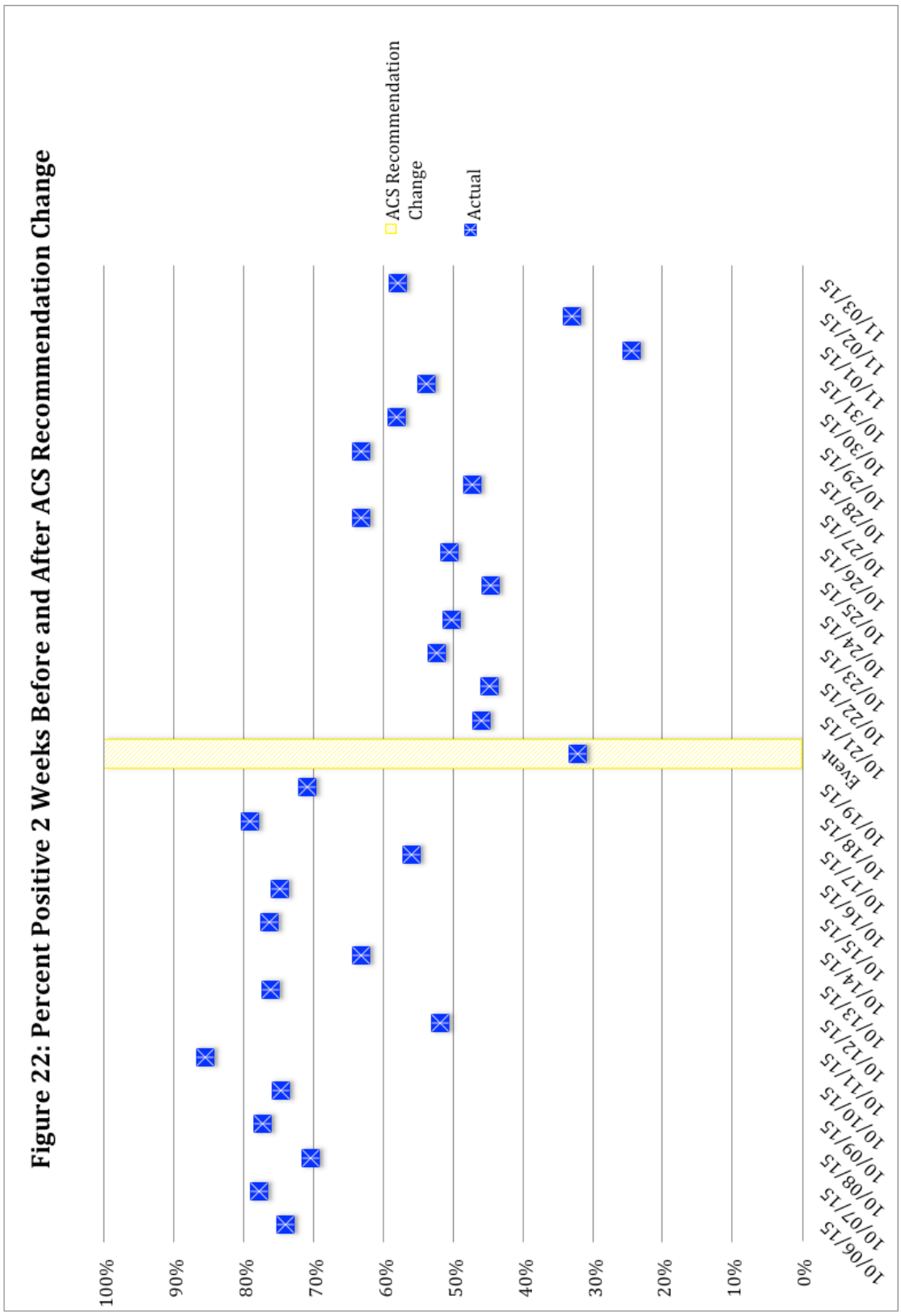


Figure 22: Percent Positive 2 Weeks Before and After ACS Recommendation Change
ACS Recommendation Change
Actual
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Event
10/06/15
10/07/15
10/08/15
10/09/15
10/10/15
10/11/15
10/12/15
10/13/15
10/14/15
10/15/15
10/16/15
10/17/15
10/18/15
10/19/15
10/21/15
10/22/15
10/23/15
10/24/15
10/25/15
10/26/15
10/27/15
10/28/15
10/29/15
10/30/15
10/31/15
11/01/15
11/02/15
11/03/15

Figure 23: Percent Neutral 2 Weeks Before and After ACS Recommendation Change

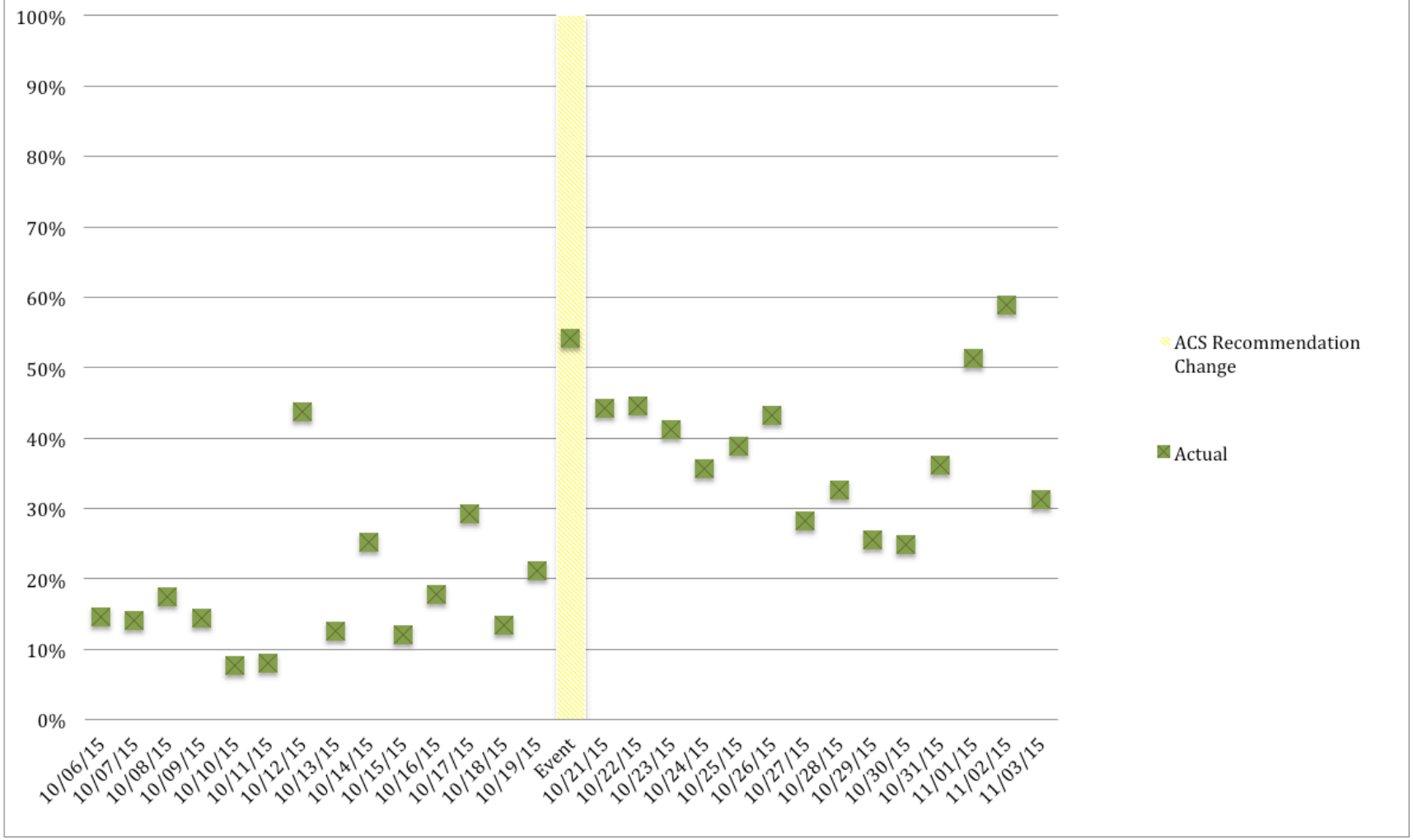
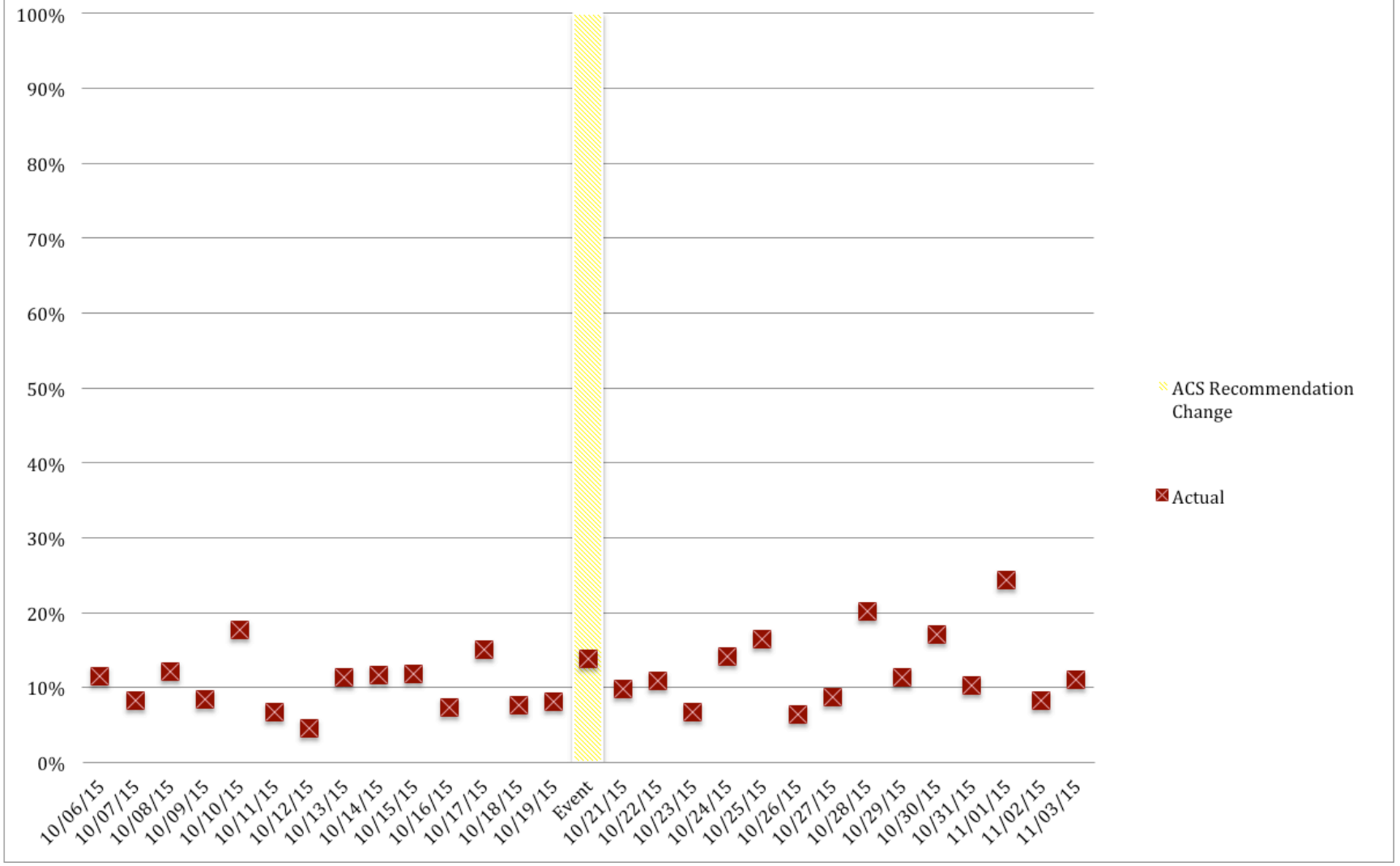


Figure 24: Percent Negative 2 Weeks Before and After ACS Recommendation Change



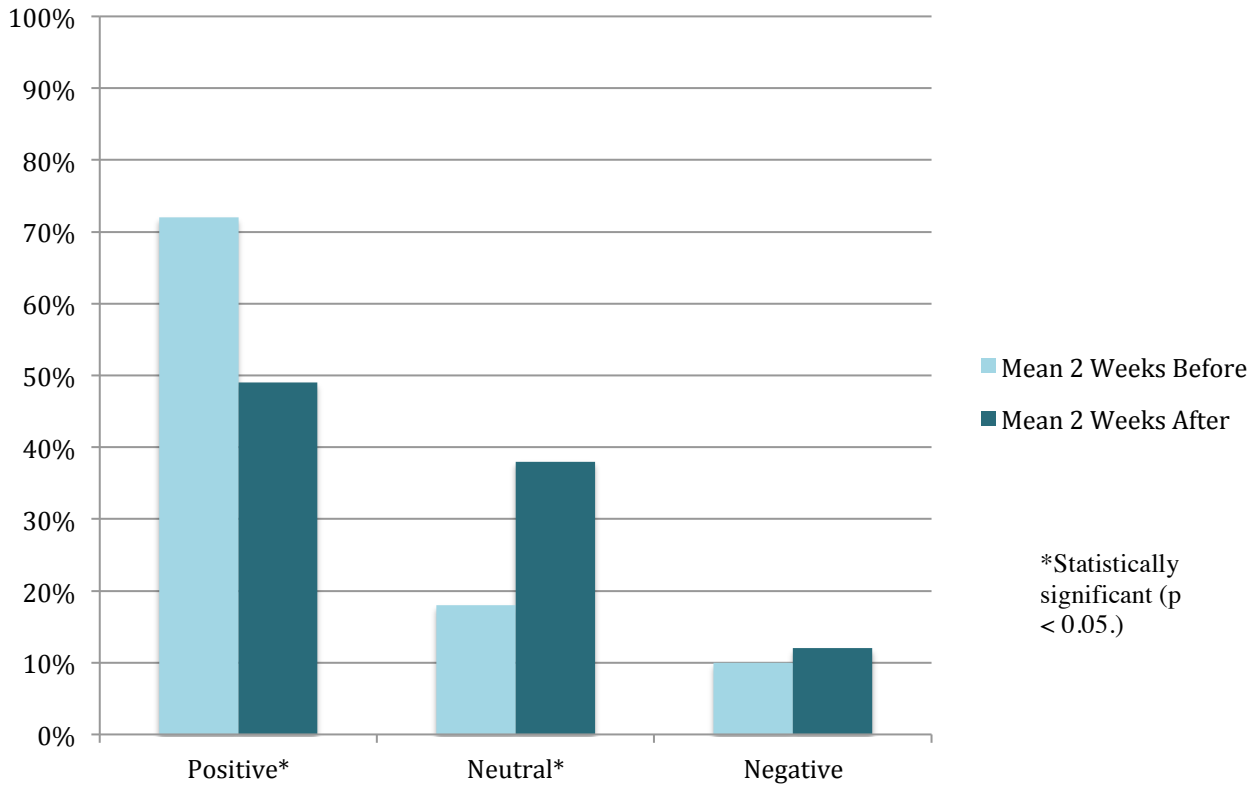
Our findings indicate that on the day of the ACS recommendation change, the Twitter dialogue experienced a decrease in positivity (Figure 22). The mean percent positive for the two weeks before the announcement was 72%. The mean percent positive two weeks after the announcement was 49%. This change is statistically significant with p = 2.3677E-06. Thus, the two weeks after the recommendation are significantly less positive than the two weeks prior. We expect that the significant decrease occurred due to the recommendation change.

The decrease in positivity suggests that either or both neutral or negative percentages experienced an increase. Figure 23 shows that during and after the ACS recommendation change neutral tweets increased. The mean percent neutral two weeks before the announcement was 18%. The mean percent neutral two weeks after the announcement was 38%. This change was statistically significant with p = 7.06E-06. The two weeks after the recommendation change were significantly more neutral than the two weeks before.

Figure 24 shows that the recommendation change did not have a statistically significant impact on negativity ($p>0.05$). The mean percent negative two weeks before the announcement was 10%. The mean percent after the announcement was 12%. We conclude that the Twitter dialogue becomes less positive and more neutral for the two weeks following the recommendation change. We hypothesize that the increase in neutral activity is due to communicating news about the announcement rather than communicating sentiment in response to the

announcement. Figure 25 summarizes the sentiment change surrounding the recommendation change.

Figure 25: ACS Announcement Effect on Sentiment Change Results



Content Analysis Surrounding ACS Recommendation Change

In order to better understand the relationship between changes to the Twitter dialogue and the ACS recommendation change discussed in the previous section, we conducted a content analysis surrounding the event. First, we performed a top word analysis to identify how long the recommendation change dominates the Twitter dialogue. Second, we determined changes in the count of tweets per category following the recommendation change in order to examine how different types of tweets respond to a recommendation change.

We began the top word analysis by creating a list of key words that are relevant to the ACS Recommendation change, shown in Table 9:

Table 9: ACS Recommendation Change Keywords

ACS
American
Cancer
Society
40
45
55
Change
Recommendation
New
Guidelines
More
Less
Prevented
Deaths
Study

We hypothesize that if the most frequent words of a given day all are present in the keywords list, then the most influential content of that day is relevant to the recommendation change.

We developed a Python script to identify the top five most frequent used words per day. The code removed stopwords and keywords shown in "Chapter 2, Table 1: Key Words". The following words were found to be the five most frequently tweeted words on the day of the ACS Recommendation change:

"guidelines," "new," "society," "american," "women"[26]

The top words for the two weeks following the event were all present in Table 9. The first day of which the top words were not present in Table 9 was November 5th. The top words on November 5th were:

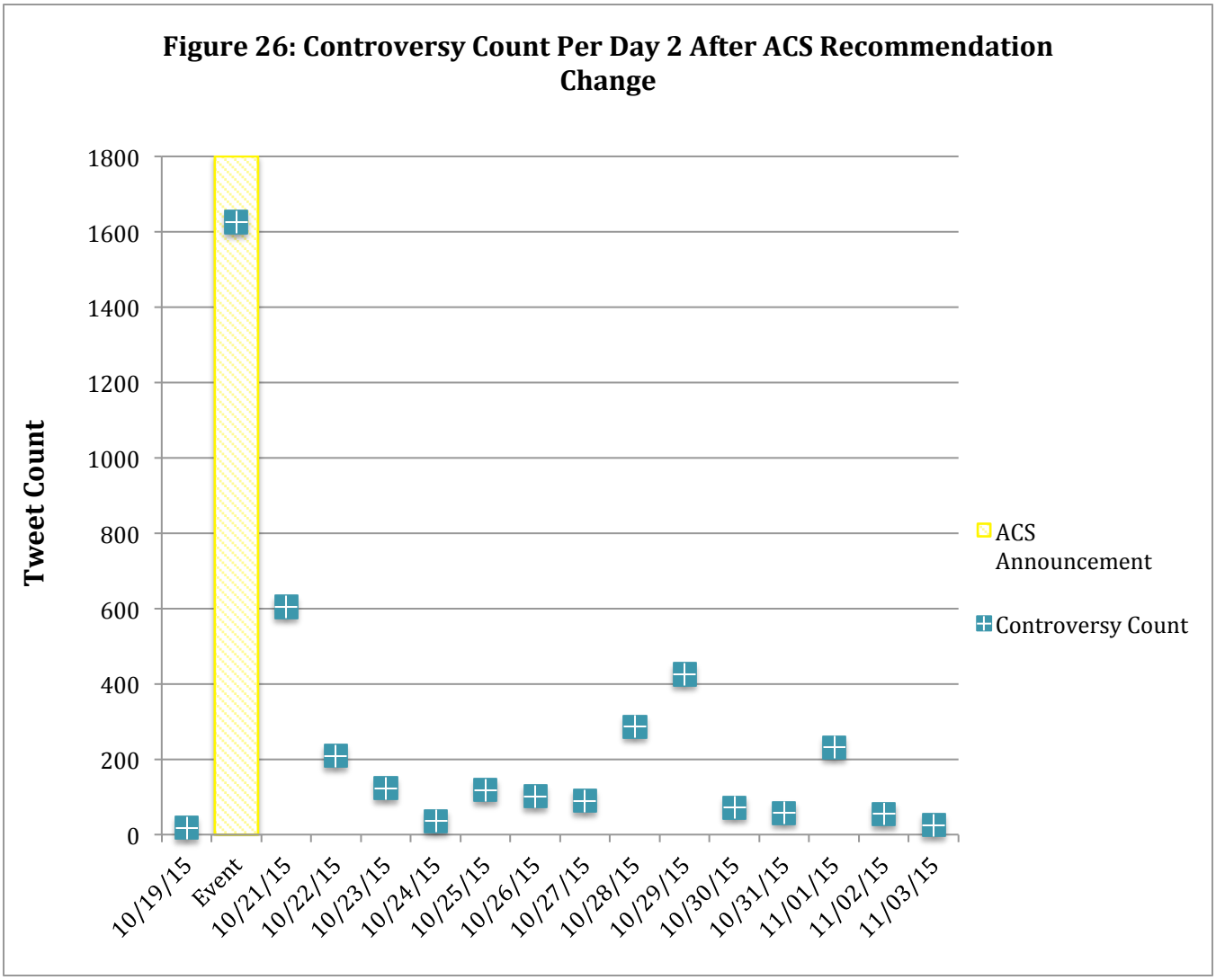
"women," "ultrasound," "adding," "results," "higher."

Thus, the Twitter dialogue is dominated by discussions pertaining to the ACS guidelines for two weeks following the event.

[26] The top words are listed most frequent to fifth most frequent.

Next, we determined the effect of the ACS recommendation change on tweet category. Due to the results of the top words analysis, we expect that controversy tweets increased for two weeks following the event. Figure 26 confirms this hypothesis: the count of controversy increases drastically the day of the recommendation change and continues to cause noise for two weeks following the event.

Figure 26: Controversy Count Per Day 2 After ACS Recommendation Change

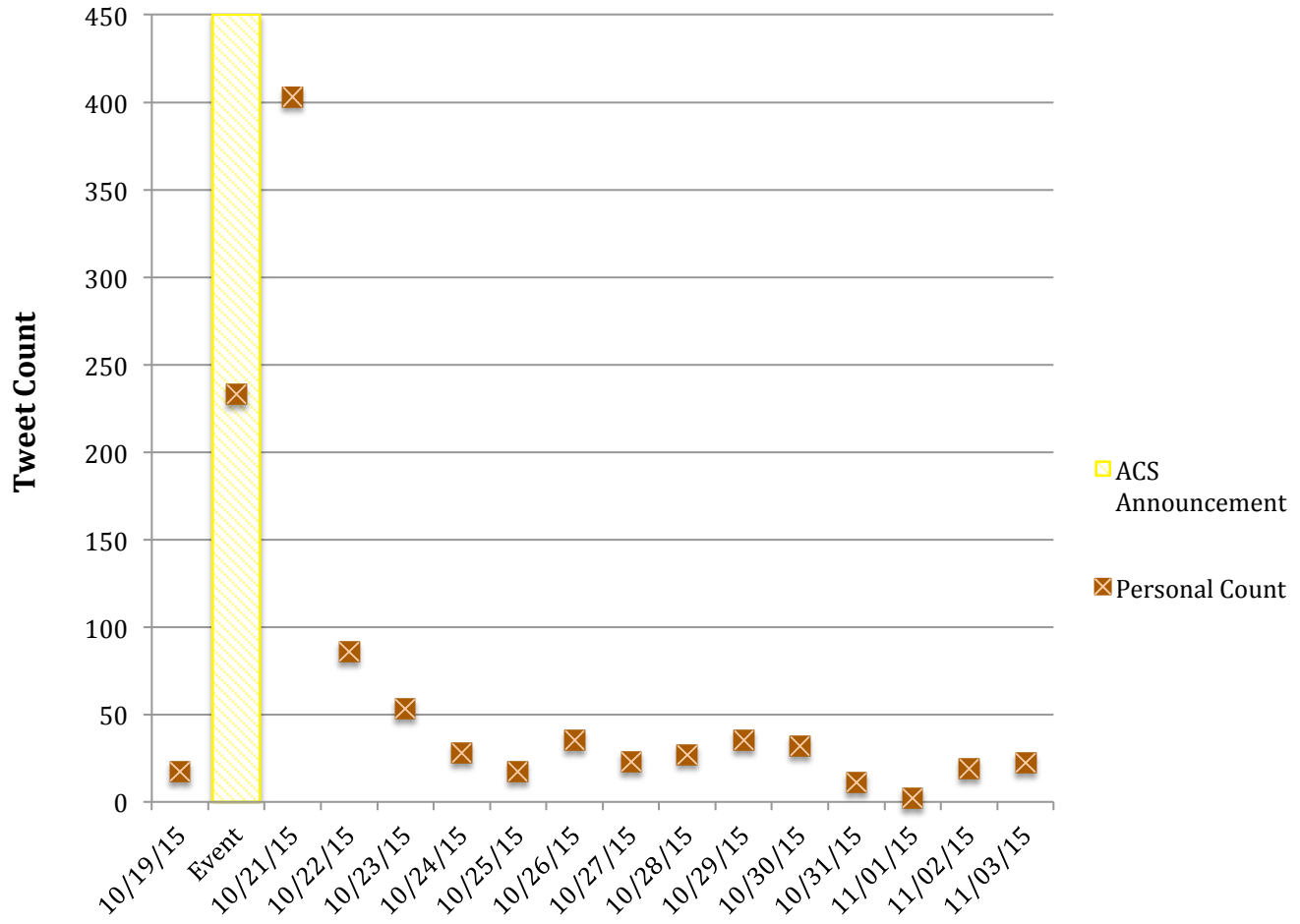


Although, advertisement and awareness spiked on the day of the event, the categories returned to normal count per day the day immediately following. We

conclude that these two categories did not experience a shift in activity following recommendation change.

Personal tweets unexpectedly spiked the day after the recommendation change. Figure 27 suggests that personal reactions to the event lag one day. The delay could be due to a delay in increased media coverage or professional communication with the public about the meaning and cause of the recommendation change.

Figure 27: Personal Count Per Day 2 Weeks Before and After ACS Recommendation



In conclusion, the ACS recommendation change impacted the Twitter dialogue after the event. Our findings suggest that the recommendation caused the Twitter dialogue to become less positive and more neutral. Additionally, Twitter content pertaining to the event dominates the dialogue for two weeks and personal tweets increase in activity following the event.

USPSTF Finalized Guideline Announcement

As mentioned in the *Introduction*, on January 11, 2016 the USPSTF finalized its decision to withhold the 2009 recommendation. The event will be considered as a two-day event: the decision was finalized on January 11th and published online January 12th. We compare the Twitter data surrounding the USPSTF decision with the data surrounding the ACS recommendation to identify if Twitter trends occur during recommendation announcements.

The findings discussed in the previous sections clearly indicate that the ACS Recommendation change impacted the Twitter dialogue pertaining to breast cancer screening. We conducted a sentiment analysis and content analysis surrounding the USPSTF announcement in order to identify if similar trends occurred in the Twitter dialogue surrounding these two events.

Figures 28 to 30 visualize the sentiment changes surrounding the USPSTF announcement.

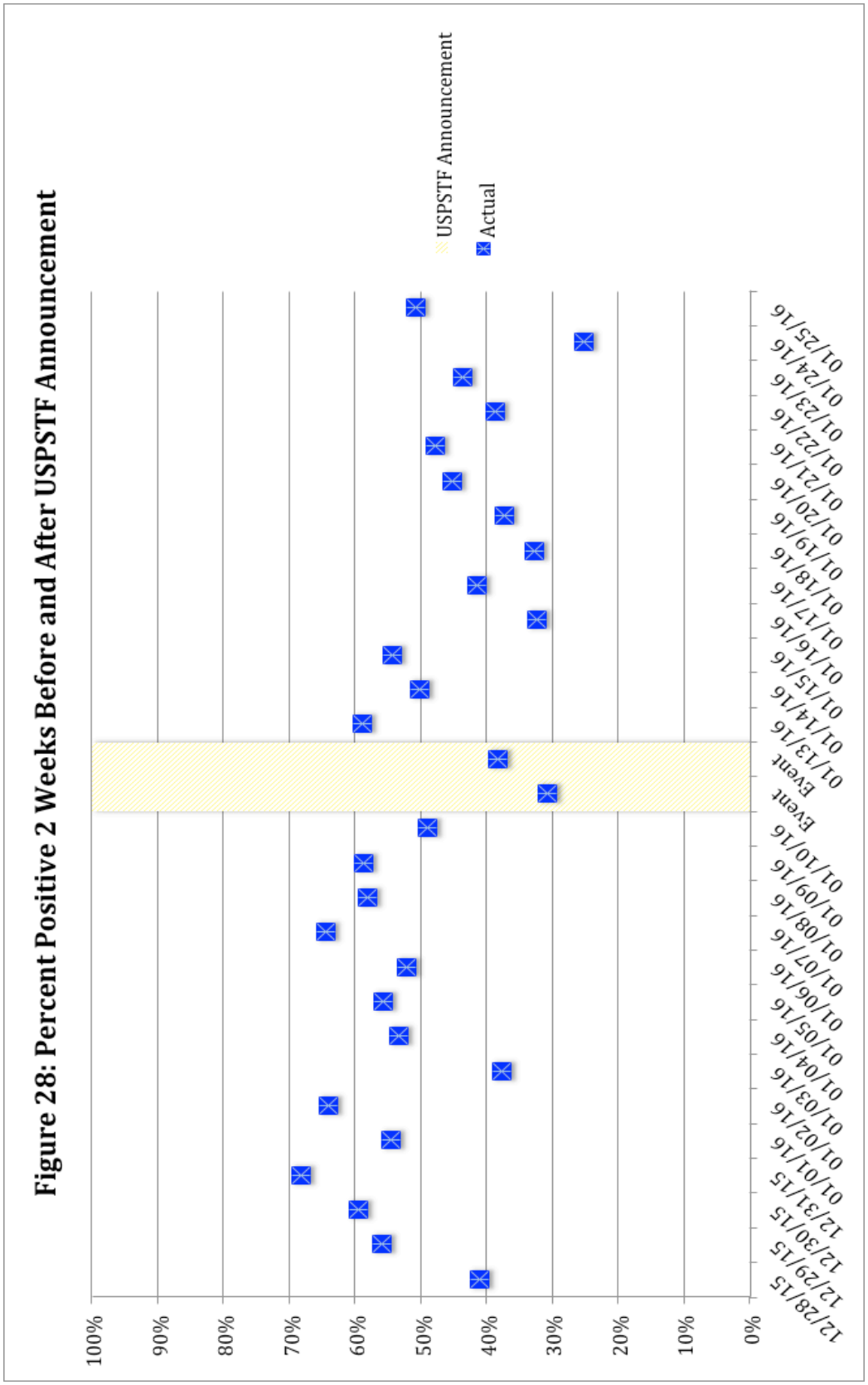


Figure 28: Percent Positive 2 Weeks Before and After USPSTF Announcement
USPSTF Announcement
Actual
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Event
12/28/15
12/29/15
12/30/15
12/31/15
01/01/16
01/02/16
01/03/16
01/04/16
01/05/16
01/06/16
01/07/16
01/08/16
01/09/16
01/10/16
01/13/16
01/14/16
01/15/16
01/16/16
01/17/16
01/18/16
01/19/16
01/20/16
01/21/16
01/22/16
01/23/16
01/24/16
01/25/16

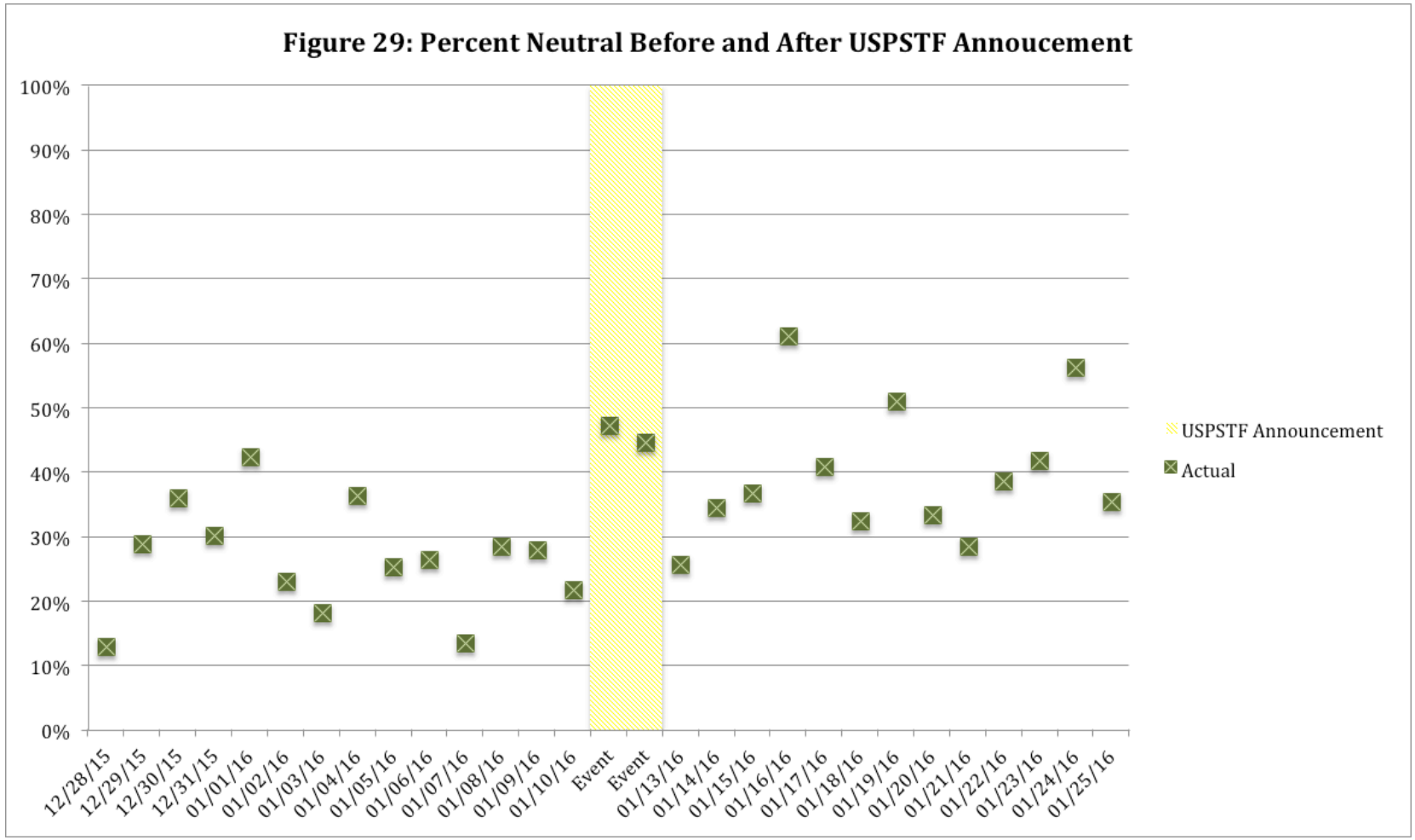


Figure 29: Percent Neutral Before and After USPSTF Annoucement
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
12/28/15
12/29/15
12/30/15
12/31/15
01/01/16
01/02/16
01/03/16
01/04/16
01/05/16
01/06/16
01/07/16
01/08/16
01/09/16
01/10/16
Event
Event
01/13/16
01/14/16
01/15/16
01/16/16
01/17/16
01/18/16
01/19/16
01/20/16
01/21/16
01/22/16
01/23/16
01/24/16
01/25/16
USPSTF Announcement
Actual

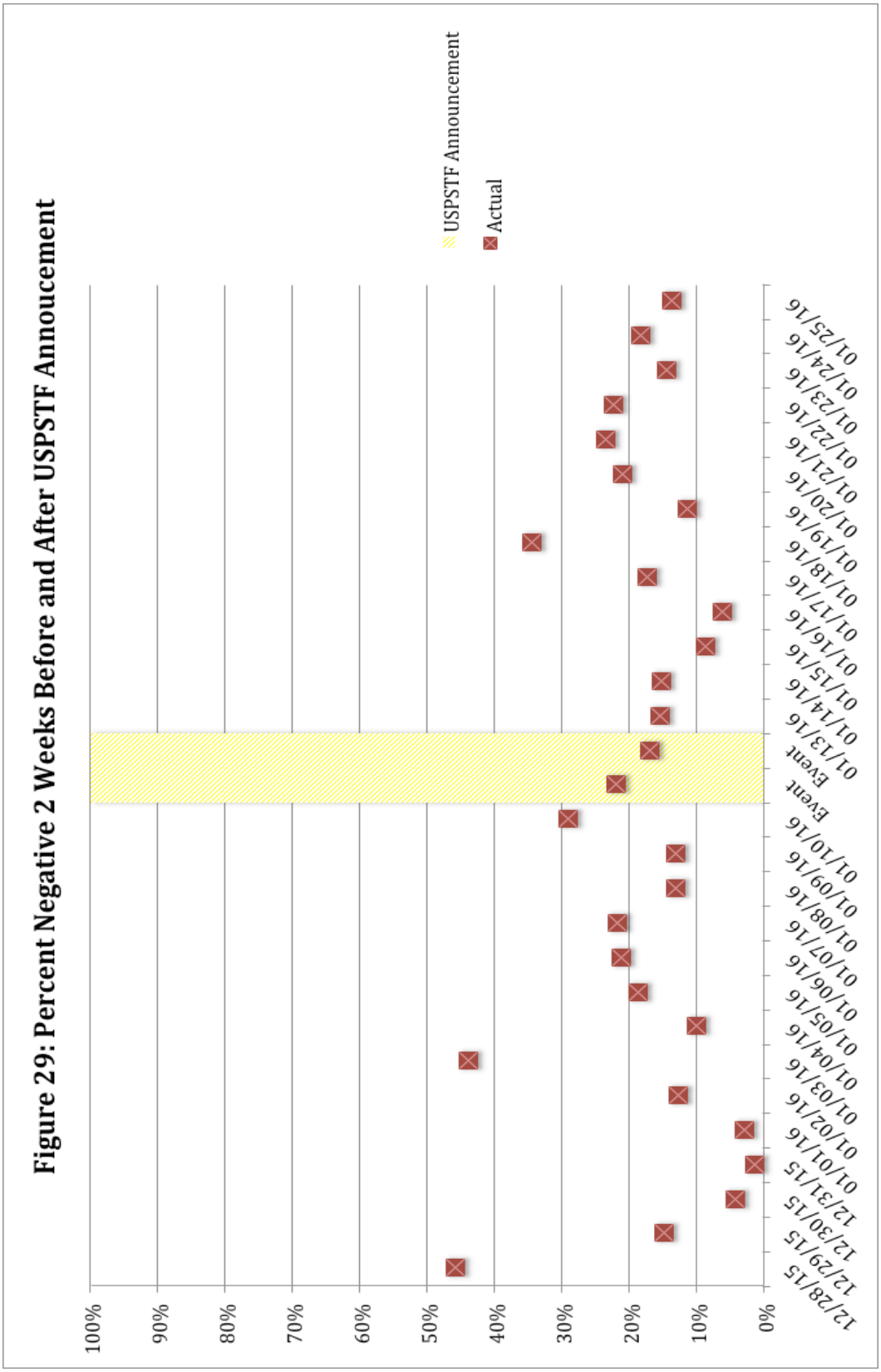
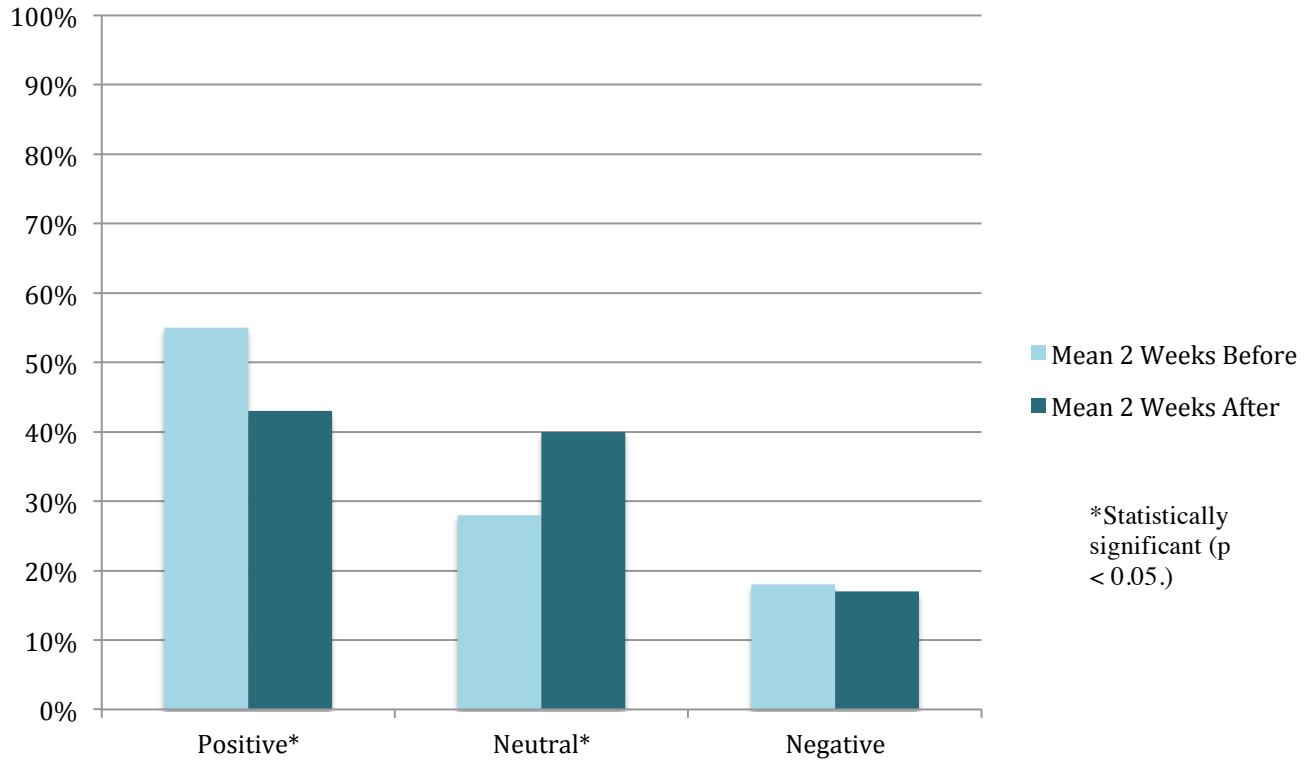


Figure 29: Percent Negative 2 Weeks Before and After USPSTF Annoucement
USPSTF Announcement
Actual
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Event
12/28/15
12/29/15
12/30/15
12/31/15
01/01/16
01/02/16
01/03/16
01/04/16
01/05/16
01/06/16
01/07/16
01/08/16
01/09/16
01/10/16
01/13/16
01/14/16
01/15/16
01/16/16
01/17/16
01/18/16
01/19/16
01/20/16
01/21/16
01/22/16
01/23/16
01/24/16
01/25/16

Figure 31 summarizes the following results: The mean percent positive two weeks before the USPSTF announcement was 55%. The mean percent positive following the USPSTF announcement decreased to 43% in the two weeks following the event. This change is statistically significant with p = 0.00165. The mean percent neutral increased from 28% to 40%. The neutral change is statistically significant with p = 0.00135. And the mean percent negative had no significant change statistically: the mean had only 1% difference between two weeks before and two weeks after the event.

Figure 31: USPSTF Announcement Effect on Sentiment Change Results



The top words analysis surrounding the USPSTF announcement was conducted using the key word list shown in table 10.

Table 10: USPSTF Announcement Keywords

Recommendation
USPSTF
Finalize
Guidelines
50
Women
New
Recommendation
Task
Force
Panel
Age

The top five words on January 11th were the following:

"recommendation," "panel," "finalize," "guidelines," "new"

The top words of each day were present in Table 10 until January 26th. The top words on January 26th were:

"pressure," "sensor," "flexible," "boost," "breastcancer"

Similar to the changes in content surrounding the ACS announcement, controversy and personal tweets experienced an increase in activity following the USPSTF announcement while advertisement and awareness did not. The results for personal and controversy activity are shown in Figures 32 and 33.

Figure 32: Personal 2 Weeks Before and After USPSTF Announcement

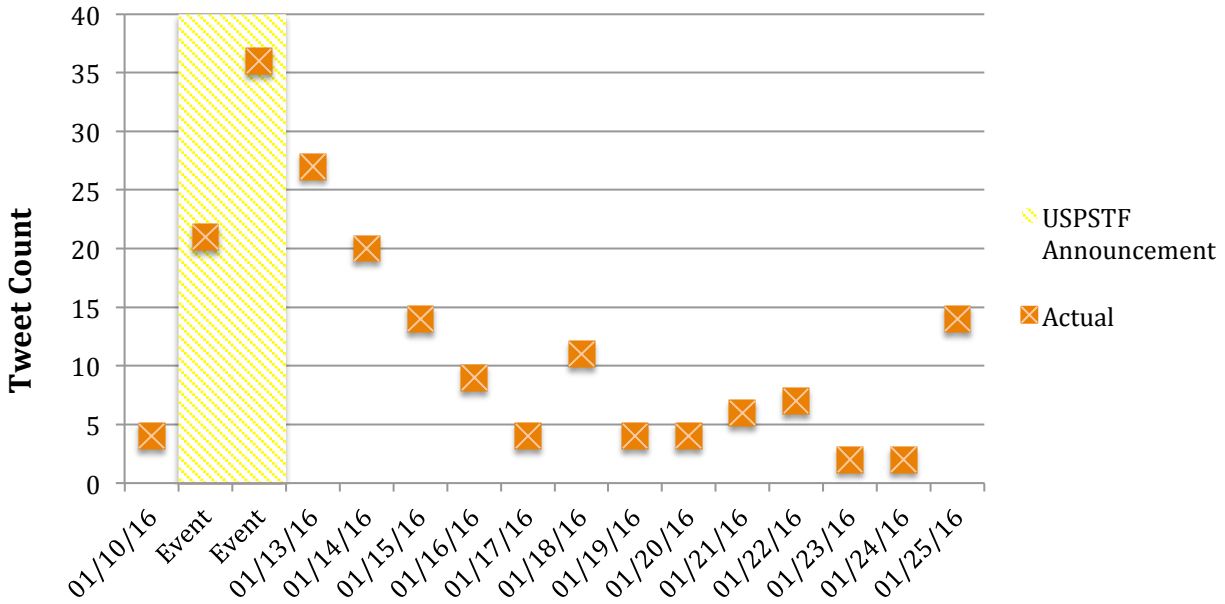
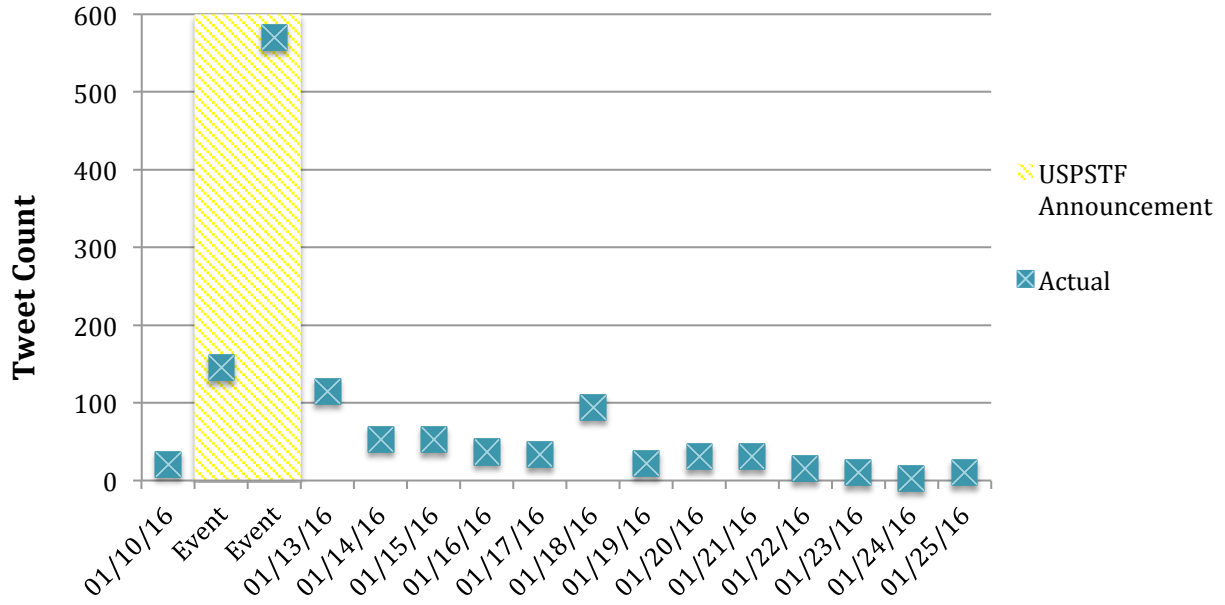


Figure 33: Controversy 2 Weeks Before and After USPSTF Announcement



In conclusion, the data surrounding the ACS recommendation change and USPSTF announcement suggests that specific Twitter trends occur surrounding events. A summary of our observations surrounding the October 20th ACS announcement is shown in column 1 of Table 9. Column 2 summarizes the observations from data surrounding the USPSTF January 11th announcement. We expect that these trends identified in the Twitter data surrounding the two events would occur in Twitter data surrounding other events related to guideline announcements as well. Future work could investigate additional events to further support our conclusion.

Table 11: Comparison of ACS and USPSTF Twitter Response

Observation	**ACS Announcement**	**USPSTF Announcement**
Volume	+5000 Tweets day	+1764 Tweets
Positivity	Decrease	Decrease
Neutral	Increase	Increase
Negative	No Significant Change	No Significant Change
Top Words	Relevant to event for 2 weeks	Relevant to event for 2 weeks
Personal	Maximum day following event	Maximum second day of event

Conclusion

Findings and Future Work

The results of this study indicate that Twitter is a useful data source for learning about public sentiment on healthcare technologies. In Chapter 1, we provided a historical context of the power Twitter data has in public health management and explored its potential use. In Chapter 2, we succeeded in data-mining almost 80,000 tweets related to breast cancer screening, trained machine learning classifiers, which classify relevance, sentiment, and category at an high accuracy of 88% to 97%, and applied classifiers to our dataset. In Chapter 3, we determined that BCAM and ACS and USPSTF guideline announcements, which occurred during the time of our study, had an impact on Twitter content and sentiment: BCAM caused an increase in tweet count in all categories as well as an increase in positivity and decrease in negativity. Both the October 20th ACS recommendation change and the January 11th USPSTF guideline announcement caused Twitter sentiment to become less positive and more neutral for two weeks. In addition, personal tweets increased in activity following both announcements. We concluded that these trends suggest specific responses to recommendation changes in the Twitter community relevant to breast cancer screening.

As mentioned in Chapter 1, this project contributes to a new and growing body of social media and machine-learning research. Unlike past studies, which used Twitter data over the course of a few weeks, the Twitter data used in our project is from the course of a full year. This fact is especially important because the timing of our data collection is what enabled us to make such a detailed analysis surrounding

recommendation changes. Previous studies did not contain recommendations from two influential organizations, the USPSTF and ACS. Furthermore, previous event analyses that focused on Twitter response to BCAM and recommendation changes did not have adequate baseline data.

Another value to our research, which was neglected from previous studies, was our incorporation of visual analysis. Our visualization includes information about the Twitter content and sentiment over time. Our interactive visualization not only allows users to understand the value of Twitter data surrounding recommendation changes, but also allows for a user to identify otherwise unknown events that gained significant Twitter coverage. The interactive visualization also includes data that is not relevant to the observed topic as a baseline. Previous research in sentiment analysis applied to specific healthcare topics did not incorporate baseline data specific to healthcare. This baseline data supports our case that the trends found in the Twitter discussion pertaining to breast cancer screening are unique and valuable.

Future work could use our analysis to determine if the trends of the Twitter dialogue correlate with trends in human behavior. One way to examine this would be to observe mammography use over the course of the past year. If certain trends occur in mammography use, we could improve the understanding of how Twitter trends and behavior are connected. Future work should also use our methodology to investigate other controversial healthcare topics during a time of policy changes.

Limitations

The following were limitations to the study:

- We collected our data through the Twitter Search API, which on days of high volume may have limited our ability to collect all relevant tweets.

- Our keyword list was determined through detailed research yet is also subject to being a biased list of relevant terms. The keyword terms were determined in the summer of 2014 and data collection continued for almost 2 years following. For the sake of consistency, we did not change the keyword list. However, a more relevant list was possible given the two year difference.

- The machine-learning process requires manual annotation, which is subject to human error. On the flip side, the machine-learning classifiers are subject to machine error.

- The data analysis was limited by the fact that we only had data from one BCAM, which was only 4 data points per day of the week.

References

1. Bifet, A., and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. *Proc. of 13th International Conference on Discovery Science*.

2. Kumar, A., & Sebastian, T.M. (2012). Sentiment analysis on Twitter. *International Journal of Computer Science Issues*, 9(4), article no 3.

3. Luciano Barbosa , Junlan Feng. (2010). Robust sentiment detection on Twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, p. 36-44*

4. Umadevi, V. (2014). Sentiment analysis using Weka. *International Journal of Engineering Trends and Technology,* 18(4).

5. Neethu, M.S., Rajasree, R. (2013). Sentiment analysis in twitter using machine-learning techniques. *Proceeding of the Fourth International Conference on Computing, Communications and Networking Technologies*, pp. 1–5.

6. Alexander Pak and Patrick Paroubek. (2010). Twitter as a corpus for sentiment analysis and opinion-mining. *Proceedings of LREC*.

7. Liu, B. Sentiment analysis and opinion-mining. *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2012.

8. Rozenblum R, Bates DW. (2013) Patient-centred healthcare, social media and the internet: the perfect storm? BMJ Qual Saf.

9. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *J Med Internet Res 2013*;15(11): e239

10. Chew C, Eysenbach G (2010) Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE 5(11): e14118.

11. Scanfeld D, Scanfeld V, Larsen EL (2010) Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control* 38: 182-188.

12. Bosley JC, Zhao NW, Hill S, Shofer FS, Asch DA, et al. (2012) Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. *Resusictation*.

13. Zhu SH, Myslín M, Chapman W, Conway M (2013) Using Twitter to examine smoking behavior and perception of emerging tobacco products. J Med Internet Res 15(8): e174. doi:10.2196/jmir.2534.

14. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: A good place to detect health conditions. *PLoS One* 9(1): e86191

15. Prabhu V, Lee T, Loeb S et al. (2014). Twitter response to the United States Preventive Services Task Force Recommendations against screening with prostate-specific antigen. *BJU Int*.

16. Sadilek, A., & Kautz, H. (2013, February). Modeling the impact of lifestyle on health at scale. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 637-646). ACM.

17. Meda, C., Bisio, F., Gastaldo, P., & Zunino, R. (2014, October). A machine-learning approach for Twitter spammers detection. In *Security Technology (ICCST), 2014 International Carnahan Conference on* (pp. 1-6). IEEE.

18. "Microblogging." techterms.com. March 2014.

19. Elmore JG, Etzioni R. Effect of Screening Mammography on Cancer Incidence and Mortality. *JAMA Intern Med.* 2015;175(9):1490-1491. doi:10.1001/jamainternmed.2015.3056.

20. Harding C, Pompei F, Burmistrov D, Welch HG, Abebe R, Wilson R. Breast cancer screening, incidence, and mortality across US counties [published online July 6, 2015]. *JAMA Intern Med*. doi:10.1001/jamainternmed.2015.3043.

21. *Help Patients Confused by ACS' Breast Cancer Screening Guidance, USPSTF, AAFP Recommend Mammography for Older Patient Group.* American Academy of Family Physicians. October 2015.

22. Lyles CR, López A, Pasick R, et al. 5 Mins of Uncomfyness Is Better than Dealing with Cancer 4 a Lifetime: an Exploratory Qualitative Analysis of Cervical and Breast Cancer Screening Dialogue on Twitter. *J Cancer Educ* 2012;28:1–7

23. *Final Update Summary: Breast Cancer: Screening*. U.S. Preventive Services Task Force. January 2016.

24. *American Cancer Society recommendations for early breast cancer detection in women without breast symptoms*. American Cancer Society. October 2015.

25. Mulcahy, Nick. *Swiss Medical Board: Stop Widespread Mammography Screening.* Medscape. April 2014.

26. Perrin, Andrew. Duggan, Maeve. *Americans' Internet Access 2000-2015*. Pew Research Center. June 2015.

27. Lam, Andrew. From Arab Spring to Autumn Rage: The Dark Power of Social Media. The World Post. November 2012.

28. *History of ACS Recommendations for the Early Detection of Cancer in People Without Symptoms.* American Cancer Society. October 2015. http://www.cancer.org/healthy/findcancerearly/cancerscreeningguidelines/chronological-history-of-acs-recommendations

29. *Final Update Summary: Breast Cancer: Screening*. U.S. Preventive Services Task Force. January 2016. http://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening

30. https://dev.twitter.com

31. "SQLite." Wikipedia. Accessed April 2016.

32. "Cron." Wikipedia. Accessed April 2016.

33. Thackeray, Rosemary. Burton, Scott. Giraud-Carrier, Christophe. Rollins, Stephen. Draper, Catherine. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC Cancer 2013*. 13:508.

34. *National Breast Cancer Awareness Month, Celebrating 25 Years of Awareness, Education, and Empowerment.* www.nbcam.org. https://web.archive.org/web/20110716123431/http://www.nbcam.org/about_faq.cfm