Obama for America 2012: A Sentiment Analysis

Nivriti Chowdhry, Matt Delhey, Jiandi Mo

October 30, 2012

Contents

1 Introduction

Utilizing the "Message Machine"[1] data set from ProPublica, a non-profit investigative journalist newsgroup, we analyzed the rhetoric employed by the Obama for America campaign through sentiment analysis of the campaign's external email communications. The original data set consisted of over three thousand unique emails collected from various additional political campaigns, including Romney for President, Newt Gingrich, and Ron Paul. The data was collected through a convenient sample: users volunteered to forward political emails they received to ProPublica, who then organized and published the data. It is also worth noting that the original intent of ProPublica was to analyze the email data with a focus on discerning how each of the campaigns is targeting specific recipients, for which they provided a web interface for comparing email variations side by side. Despite this focus, there is more than a sufficient amount of data within the

[1] More information about this data set and the ProPublica organization can be found at http://projects.propublica.org/emails/

emails and users databases to provide insights on the type of commutative of language utilized by the Obama campaign.

Before exploring the emails and conducting any sentiment analysis, we first cleaned the data by subsetting for the Obama for America campaign. One motivation for focusing our topic solely on the Obama campaign was the fact that it was responsible for over 68% of the registered emails, with only less than 9% of the registered emails in the data set coming from the Romney for President campaign. Next, we chose only to consider emails past the date 2012/04/30 because no day prior had received anything but a trivial number (less than ten) of registered emails. Finally, we created emotion and polarity variables using the sentiment analysis methodology fully described in the next section.

Our first two findings relate to the idea of change in sentiment over time. We begin by first analyzing the trends in the change in sentiment from a general perspective. We then introduce six interesting political events and superimpose these over our change over time plots. We argue that, given the significance of the events and the corresponding observable change in sentiment, these events likely have causal relationship with the change in sentiment. Employing this premise, we conclude our first finding by conducting a sentiment analysis of the Obama for America campaign for one particularly correlational event.

For our second finding, we continue this form of analysis to the remaining five political events and then conclude by making some general observations about the combinations of sentiments used by Obama, such as tendency to combine fearful and neutral language as well as surprising and positive language.

To reach our third finding, we used a comparison of the sentiments in President Obama's Twitter Feed and the Obama for America campaign emails. The comparison shows that there are differences in the role that each plays in the campaign. The Obama for America campaign emails serve as a tool to gather support for President Obama. They interact with existing supporters while also introducing potential and new supporters to President Obama's ideas and values for the second term of his presidency. President Obama's Twitter Feed serves a more personal role as a tool public interaction; the Twitter account makes President Obama more personable. The sentiments between the two tools are usually similar, but differ in observable and interesting ways. The greatest difference was between the sentiments of joy and surprise in President Obama's tweets after the Supreme Court decided to uphold the Patient Protection and Affordable Care Act, possibly because President Obama was pleased with the outcome. The emails from the Obama for America campaign were more disgusted in sentiment, likely because they were addressing the situation in a more abstract sense and thus also including the concerns of disgusted supporters who were negatively impacted by the Supreme Court's decision. President Obama's Twitter Feed also shows small peaks in anger at major events that we explored.

Finally, for our fourth finding we conducted a case study of the Patient Protection and Affordable Care Act and found that sentiments conveyed in the emails circulated by the Obama for America campaign varied around the date that the Supreme Court decided to uphold the Patient Protection and Affordable Care Act. Prior to the decision on June 28, 2012, the Obama for America emails were marked with sadness, possibly to target email recipients' empathy and convince them to support the Patient Protection and Affordable Care Act. Following the Supreme Court's decision, the greatest spike in sentiment was that of disgust. However, only two of six income brackets showed a spike in disgust after the Supreme Court's decision, namely annual household incomes of less than $35,000 or more than $150,000. Though the emails do not convey disgust towards the recipients, they do address the concerns of the disgusted recipients. Families with annual household incomes lower than $35,000 were likely upset about having to pay a penalty for not having medical health insurance while families with annual household incomes greater than $150,000 were likely upset about having to pay higher and more taxes to fund the federal subsidies families with annual household incomes of less than $88,200.

2 Explanation and Details of Our Sentiment Analysis

The sentiment analysis was conducted in R by using the `sentiment` package[2] available on the CRAN repository. The `sentiment` package provided us with two distinct functions for conducting our sentiment analysis: `classify_emotion` and `classify_polarity`. These functions operate by taking the input of our data frame of emails and outputting additional sentiment classifications for each row or email. Both were utilized in producing our overall sentiment analysis.

More specifically, the `classify_emotion` function attempts to classify each email by emotion, operationally defined as anger, disgust, fear, joy, sadness, and surprise, by implementing naive Bayes classifiers trained on Carlo Strapparava and Alessandro Valitutti's emotions lexicon. This outputs seven numerical values representing the absolute log likelihood of the email expressing each given emotion and a categorical value indicating the most likely sentiment category for each email or, if not enough information is available, NA.

The `classify_polarity`, operates in a similar fashion for classifying polarity utilizing a naive Bayes classifier trained on Janyce Wiebe's subjectivity lexicon. This function outputs two numerical values representing the absolute log likelihood of the email expressing a positive or negative sentiment, a value representing the ratio of the absolute log likelihoods, and a categorical value indicating the most likely sentiment category for each email between between positive, neutral, and negative.

The field of sentiment analysis and applied natural language processing is clearly deeper and more complex than the treatment it has been given here. While we recognize the limitations of our sentiment analysis methodology, we still hold that the conclusions we have drawn from it are genuinely interesting and appropriate for the scope of this project. From reading exemplar emails and comparing our personal sentiment analysis to that of the algorithm, we believe that the results match our intuitions and allow us to make valid conclusions.

3 Sentiment Over Time: General Overview & Supreme Court Upholds Obamacare

3.1 General Overview

The presidential election cycle is perhaps best known for its volatility. During this time period, the presidential candidates and the rest of the political world are simply in a period of flux. With this understanding, it seems reasonable to wonder if the sentiment of the language used by politicians to communicate with the public is in flux as well. In an attempt to answer this question for the Obama campaign, we looked at the change over time in the emails for the classifications of both polarity and emotion. While we are only looking at emails within a time span of about six months, during a competitive election cycle such as the one we are currently experiencing six months is ample time for each candidate to experience the political highs and lows that should bring about observable changes in the sentiment of the language used, and this is what we find reflected in the data for the Obama campaign.

In the following plots found in 1 we have presented two perspectives on the change in sentiment: the change in emotion and the change in polarity. In order to provide some context and explanatory force to the plots, we have also selected and superimposed six important political events. The events that we have chosen, the date that they occurred, and their label in the plots can be found in the table below:

From the change in polarity over time plot found at the top of 1 we can notice that the best fit polarity is rather volatile over time. The general trend of the plot shows us that the emails received from the Obama

[2]Credit to the original author Timothy P. Jurka. More information on the sentiment package can be found at http://cran.r-project.org/web/packages/sentiment/sentiment.pdf

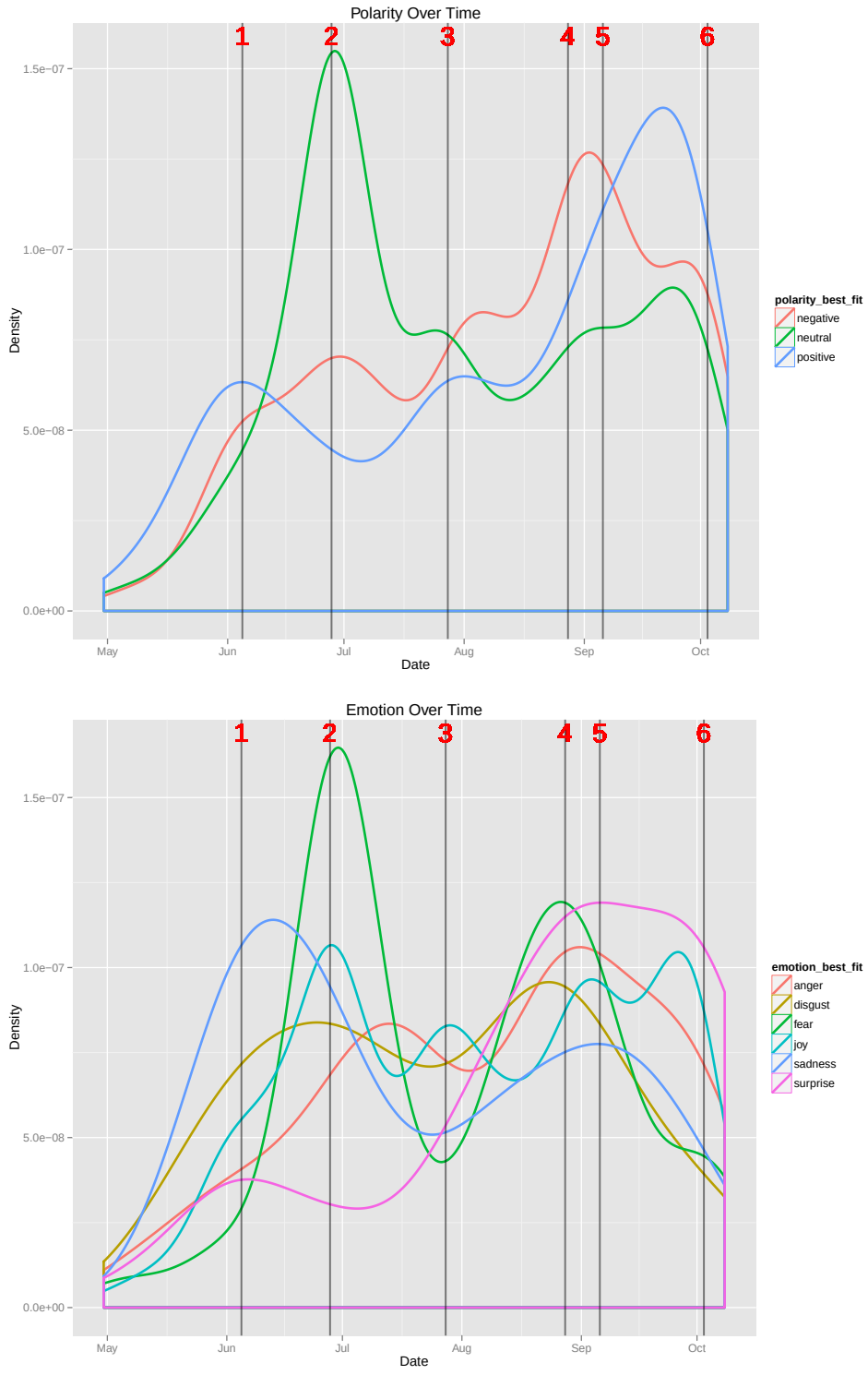

Figure 1: (Top) A density plot displaying the change in the polarity of emails over time, with selected significant political events superimposed. Notice the early peaks of positive and neutral polarity corresponding to events one and two. (Bottom) A density plot displaying the change in the emotion of emails over time, with the same selected significant political events superimposed. Notice how the peak in fear corresponds to the peak in neutral polarity.

Label	Date (2012)	Political Event
1	6/05	Primaries in 6 states
2	6/28	Supreme Court upholds Obamacare
3	7/28	FEC fundraising deadline for the Fall
4	8/28	Republican National Convention
5	9/06	Democratic National Convention
6	10/03	First Presidential Debate

campaign tended to be neutral during June and July and then later tended to be negative throughout August and positive throughout October. From the change in emotion over time plot found at the bottom of 1 we can notice that the best fit emotion of the emails is similarly volatile over time. One feature of the plot that is particularly interesting is the influx in the density of fear that corresponds to the influx in the density of neutrality, implying a possible correlation between the two sentiments.

3.2 Supreme Court Upholds Obamacare

But what are the relationships between these events and the data? While the changes over time in email sentiment may be interesting by themselves, without any context of what the Obama campaign might have been communicating with the public at the time of their delivery they seem to both lack a sense of substance and occur without any real cause.

The first and most correlated political event with our changes over time is the Supreme Court ruling in favor of Obamacare. The fact that this ruling, perhaps one of the most significant in Obama's presidential legacy so far, inspired an observable change in sentiment is no surprise. What is surprising is that the ruling caused by the far the greatest change in both polarity and emotion. Even more interesting is just what specific polarity and emotion best fits our emails corresponding to this event, namely the combination of neutral polarity and emotional fear. The data gives us evidence that after the ruling rather than celebrating and gloating through email with his supporters, Obama instead addressed them with a level head and a cautious perception of the future. One objection to this inference might be that our sentiment analysis algorithm has made a mistake. While this is admittedly a possibility, it seems unlikely that we would *not* see some kind of divergent sentiments in the emails during this time period. Given this observation our conclusion regarding Obama's expressed sentiments seems to be a strong possibility even in the face of hostile skepticism.

4 Sentiment Over Time: The Remaining Events

4.1 Primaries in 6 states

We can continue this analysis onto the other political events, starting with the primaries in California, Montana, New Jersey, New Mexico, South Dakota, and North Dakota[3]. This is the first date within our time-frame that we see a significant change in the rhetoric used by the Obama campaign in their emails. Looking at the polarity change over time plot at the time of these primaries we see a peak in the positive sentiment relative to the other polarities. A similar peak is reflected in emotion best fit plot, in which we see a beginning of a peak in the emotion sadness. The data seems to say that the campaign communicated sadly with a positive polarity. Given the context that the Obama campaign was already assured of earning the nomination from the Democratic Party, this conclusions from the data seems plausible.

[3]North Dakota technically holds a Democratic caucus.

4.2 FEC fundraising deadline for the Fall

The FEC fundraising deadline is of obvious importance for the Obama campaign. It's their last chance to gather general fundraising money for some spending means. In order to make this last push, the polarity over time plot seems to tell us that the campaign used a roughly equal mixture of all three polarity sentiments. The emotion over time plot tells a similar story, indicating a cross between multiple emotions. Using our understanding of fundraising and its importance to any presidential campaign, we can conclude from the data that most likely the Obama for America campaign was attempting to push everyone's buttons, as it were, by utilizing a diverse set of sentiments.

4.3 Republican National Convention

In American politics, it is generally the rule that a political party during their opponent's national convention stays relatively quiet in terms of pushing new policies. That is, they give their opponents the policy making spotlight for a week and either patiently wait for their own national convention in the future or continue to analyze the results of their own recent convention. For 2012, Obama fell within the first category and therefore his campaign had a preparatory Republican National Convention. The data seems to tell us that the Obama for America campaign might have defied this general policy by communicating emails that consisted somewhat densely of negative polarity emails along with a moderate density of emotive fear.

4.4 Democratic National Convention

The Democratic National Convention was the Obama campaign's first chance to really generate momentum for the final leg of the electoral season. The data shows the campaign decreasing their negative and fearful sentiments while increasing their surprise, joy, and anger emotions and positive polarity of their emails. We can see that the trend for positive emails continues all the way until mid-September, when the campaign will finally be losing its convention hype and preparing for the Presidential debates. This pattern seems to be reflected in the emotive sentiment, in which the data suggests that the Obama campaign tapered off its surprise and anger emotive language in favor of an increase in fear.

4.5 First Presidential Debate

The first Presidential Debate comes at the tail end of our data set and therefore we see our density decreasing here for both polarity and emotional sentiment. What's interesting is that we can notice a similar pattern between the joy emotion sentiment and the neutral polarity sentiment. Similarly, we see the surprise emotion with the highest density during the Presidential Debate along with the positive polarity. What can we conclude about Obama's campaign rhetoric from this scenario? Not as much as the previous cases, but we can infer from the data that Obama was at the very least communicating quiet differently than he was during the summer, perhaps with a stronger focus on positive and surprising statements.

Overall, we can see that there are direct correlations between some of the emotional and polarity sentiments, specifically between those between fear and neutral and positive and surprise, and perhaps anger and negativity. While these are intuitive ideas to us through common sense, they are interesting results from the data because they don't seem like necessary means for which Obama would need to communicate with his intended audience, but instead optional and possibly casually related. Furthermore, we can conclude that the Obama campaign behaved as we might such expect, expressing various and volatile types of sentiments throughout the election cycle.

5 Sentiment Analysis & Comparison of Obama's Twitter Feed

A comparison of the sentiments in the tweets on President Barack Obama's personal Twitter account and the sentiments of the emails circulated by the Obama for America campaign highlight both the common sentiments between the two and President Obama's stability and diplomatic state over a period of six months from April 2012 to October 2012.

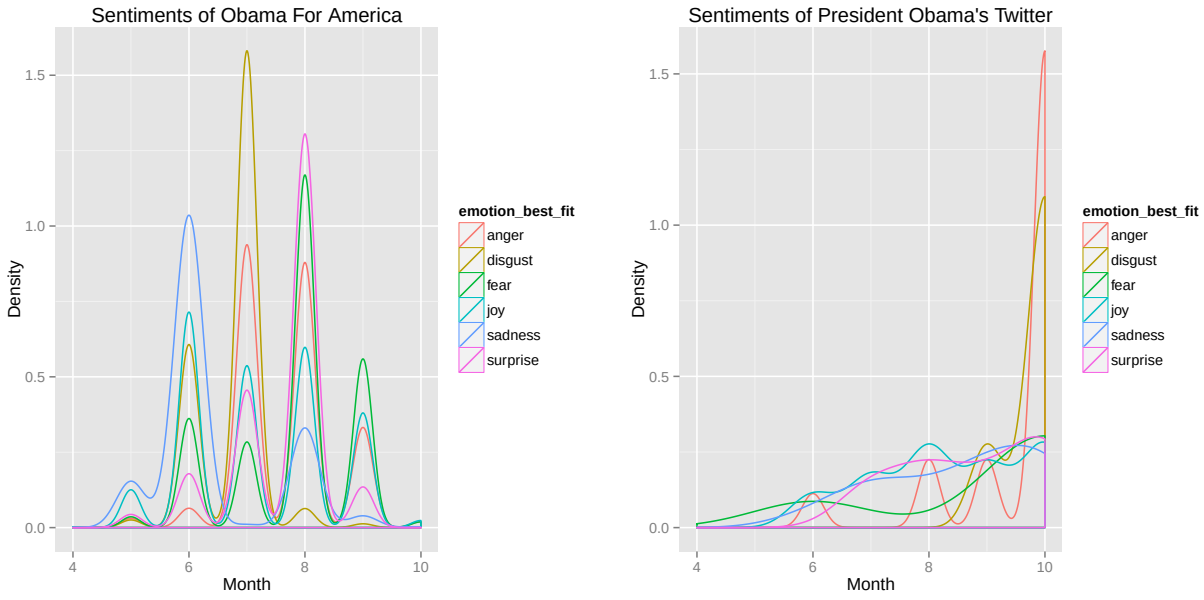

Figure 2: (Left) Density plot showing different emotions associated with the emails sent from the Obama for America campaign. (Right) Density plot showing the change in sentiments in President Obama's personal Twitter. Note: Obama displays the most sentiment towards the end of September.

As mentioned above, six key events passed during this time: preliminaries in 6 states, the Supreme Court's ruling over the Patient Protection and Affordable Care Act, FEC fundraising deadline for presidential candidates in the fall, the Republican National Convention, the Democratic National Convention, and the first Presidential Debate. Although the sentiments associated with each of these events through the emails from the Obama for America campaign are accounted in the previous section, Obama's personal sentiments, as determined through his personal Twitter account, differ.

President Obama displayed anger during the preliminaries in six states while his campaign's emails conveyed more sadness. The combination of the two indicate that President Obama may have been falling behind in the polls slightly, making President Obama frustrated and sounding angry, while the campaign emails mourned a slight loss.

When the Supreme Court decided to uphold the Patient Protection and Affordable Care Act, President Obama's Twitter was filled with joyous and surprised sentiments, while the Obama for America campaign sought to appease disgusted email recipient and address their concerns. In their effort to appease these recipients, the emails from Obama for America sounded more disgusted as well, as discussed in more detail in the case study at the end of this paper.

As the FEC deadline for the fall approached at the end of July, both President Obama's Twitter and the Obama for America campaign displayed sentiments of anger, while the Obama for America campaign also displayed surprise and fear. The sentiments of anger may have come from the stress associated with a quickly approaching deadline, while the surprise and fear found in the emails from the Obama for America campaign

could have been inserted to encourage people to donate more money to the campaign. The sentiments of fear may have come from foreboding a less appealing future due to lack of funding while surprise may have come from emails attempting to inspire people to donate more money.

The Republican National Convention was held a week before the Democratic National Convention. During the Republican National Convention, Obama displayed anger and disgust, most probably against the Republican candidate and how the Republicans represented themselves at the convention. The Obama for America Campaign displayed fear and anger, possibly in emails warning Obama's supporters of the threat the Republican party was becoming or providing information about the upcoming Democratic National Convention, during which there was a decrease in the anger and disgust conveyed in Obama's Twitter updates. The sentiments of the Obama for America campaign remained largely the same.

Lastly, as the first Presidential Debate approached at the beginning of October, President Obama's Twitter was more frequently updated while there were fewer emails sent out through the Obama for America campaign. This could be to help President Obama become a more personable figure whom people could relate to or because the FEC deadline for the fall had passed and there was no point sending more emails to solicit monetary support.

Overall, President Obama's personal Twitter shows small spikes in anger during each of the six major events listed throughout the paper. Perhaps they are tweets intended to be persuasive, inspiring, or inspirational, but that end up registering as anger. The roles that the Obama for America campaign emails and President Obama's Twitter updates play in the election are very different. The Obama for America campaign gather support for President Obama throughout the entire campaigning process. President Obama's Twitter is more of a self-promoting tool that President Obama used towards the end of the campaign as he moved into the debates. Through the early stages of the campaign, people were focused on each presidential candidates' ideas and values. Closer to the debates, people started becoming more invested in individual candidates. It became time for the candidates to promote themselves, which is where President Obama's Twitter realized a role.

6 Case Study: Patient Protection and Affordable Care Act

On June 29, 2012, the United States Supreme Court decided to uphold the Patient Protection and Affordable Care Act, under which every individual in the United States was required to have some form of medical health insurance. Individuals and their dependents who were not covered by medical health insurance policies issued by their employer, Medicare, Medicaid, or some other public health insurance plan need to purchase private medical health insurance of pay a penalty. Although the Patient Protection and Affordable Care Act requires medical health insurance, even for low income families who have never before been able to afford health insurance, it also provides federal subsidies for households with incomes lower than four times the national poverty level. Therefore, any household with an income of less than $88,200 would receive some amount of federal subsidies, while families with combined incomes below 150 percent of the poverty level, or $33,075, would only be required to pay 2 percent of their income, which comes close to $50 each month.

Since the Patient Protection and Affordable Care Act is inherently based on household income, an exploration into the sentiments associated with different income levels is warranted. However, before diving into a detailed exploration of whether emails with different sentiments are sent to different income groups and which sentiments can be attributed to each group, the data must be cleaned. As for the rest of the findings in this paper, the emails analyzed were subsetted to include only those sent from the Obama for America campaign to fix for the different opinions possibly conveyed within the emails. Any political campaign, like Obama for America, must choose its stand on a relevant issue and maintain that stance throughout the campaign. Therefore, Obama for America maintained a single opinion on the Patient Protection and Affordable Care Act, but likely conveyed that stance to different income brackets using different language

and sentiments, which is what this case study strives to discover.

Once the data is cleaned and subsetted to represent only emails sent through the Obama for America campaign, the emails were scanned for the word ”tax”. The keyword was chosen based on our intention to understand whether the additional taxes that would result from the Patient Protection and Affordable Care Act had an effect on the sentiments associated with its passing. After subsetting for both the Obama for America campaign and tax discussions within the emails, a histogram was created to show which sentiments were associated with emails as a whole.

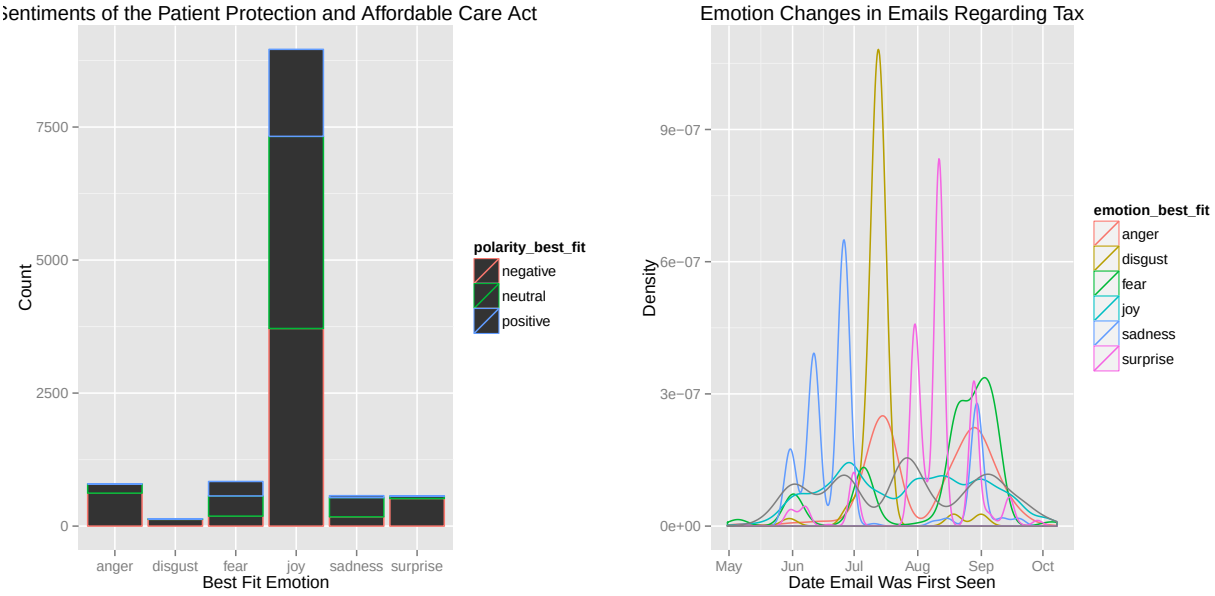

Figure 3: (Left) Histogram showing the difference in the occurrence of each sentiment in emails regarding tax. (Right) Density plot showing the change in sentiments regarding tax policy over time. Note: different colored lines represent different sentiments.

The histogram in Figure 3 pictured above shows that joy was by far the most common emotion associated with taxes, while the sections of each bar show whether they the emotion was conveyed using negative, neutral, or positive polarization. While this graph provides a general understanding of the emotions of the campaign emails, a view at how the emotions changed over time help better understand how sentiments towards taxes changed overtime. Therefore, changes in the best fit emotion over time were analyzed.

The second graph in Figure 3 pictured above shows that, although joy is known to be the most common emotion within the emails, there are large spikes in other emotions from May 2012 to October 2012. Sadness peaks at the end of June and surprise peaks towards the middle of August. The most noticeable peak, however, is that of disgust in early July. The Supreme Court decided to uphold the Patient Protection and Affordable Care Act on June 28, 2012. The high density of emails best with a sadness sentiment indicate that the emails circulated during the days leading up to the Supreme Court’s decision may have been trying to gather support for the Patient Protection and Affordable Care Act. Once the Supreme Court decided to uphold the act, emails aimed at gathering support decreased, and soon thereafter, emails with a disgusted sentiment increased.

The early part of July is the only time during with emails with a disgusted sentiment were sent through the Obama for America campaign. Since the peaks in sadness and disgust surround the dates of the Supreme Court’s decision, further speculation into variables related to the Patient Protection and Affordable Care Act could confirm the relationship of sentiment changes to the Supreme Court’s decision and the people most

affected by the decision.

As discussed earlier, the Patient Protection and Affordable Care Act is largely based on household income, because of which we looked at how different income brackets are targeted with emails using different sentiments.

The final in plot Figure 4 on page 11 graph shows that emails sent to different income groups do, in fact, convey different sentiments. Each income bracket shows a spike in disgust in early July. The poorest and richest income brackets show the greatest spikes in sentiment changes as all the disgust conveyed in emails sent to those groups of people are concentrated in early July. This indicates that the Obama for America campaign was sending these groups different messages with different sentiments than the others, possibly because these groups of people responded to the Supreme Court's decision to uphold the Patient Protection and Affordable Care Act differently than other income brackets. This could be because the Patient Protection and Affordable Care Act most deeply affected households earning less than $35,000 in annual income and households earning more than $150,000 in annual income.

Low income groups, like the families making less than $35,000 annually, may have been negatively affected by the Patient Protection and Affordable Care Act, which required the Obama for America campaign to address their disgust in their emails. It is possible that the poor were upset about having to pay a penalty for not being insured through an employer, a public medical health insurance provider, or private medical health insurance. It is possible that people with such low incomes did not pay for any kind of medical insurance, and may have felt that the required penalty that was 2% of their annual income was more than they had ever given for medical insurance before.

On the other end of the spectrum, wealthier families that earned more than $150,000 ended up bearing the cost of both the Patient Protection and Affordable Care Act and increased Medicare taxes that were associated with it. The Patient Protection and Affordable Care Act encompasses a 40% excise tax on annual health insurance premiums, meaning that families who are privately insured may end up bearing some of the excise tax. Furthermore, the Medicare tax applies only to households earning over $200,000 annually, $250,000 in some states. These families are included in the bracket of households earning over $150,000 annually, which is why the spike in disgust could be the concerns the Obama for America campaign had to address to appease the people who were upset about paying increased taxes.

7 Conclusion

The 2012 United States Presidential Election is finally wrapping up after months of steadfast campaigning from all the candidates. The focus of this paper was the Obama for America campaign and how it conducted itself and managed its rhetoric utilizing an analysis of six major events. These events marked significant stages in the election, starting with primaries in six states, the Supreme Court's decision to uphold the Patient Protection and Affordable Care Act, the FEC deadline for the fall, the Republican and Democratic National Conventions, and the first Presidential Debate. Each of these events was associated with a different set of sentiments, which are understood through a sentiment analysis of the content of the emails the Obama for America campaign sent out.

From this sentiment analysis of email content over time, we argued that the trends seen in the change in sentiment were most likely caused by these political events. Looking more closely at these trends, we found that each event had a corresponding change in sentiment in the emails from the Obama campaign that could be used to paint a general picture of how the campaign relatively responded to these events. The sentiments found in the Obama for America emails were found to be volatile, changing, and situation, much like the political campaign season itself.

These sentiments were then compared with the sentiments found in President Obama's Twitters updates. President Obama's sentiments were much more leveled than the sentiments in the Obama for America Campaign's emails. However, the anger sensed in President Obama's Twitter updates peaked during of the six

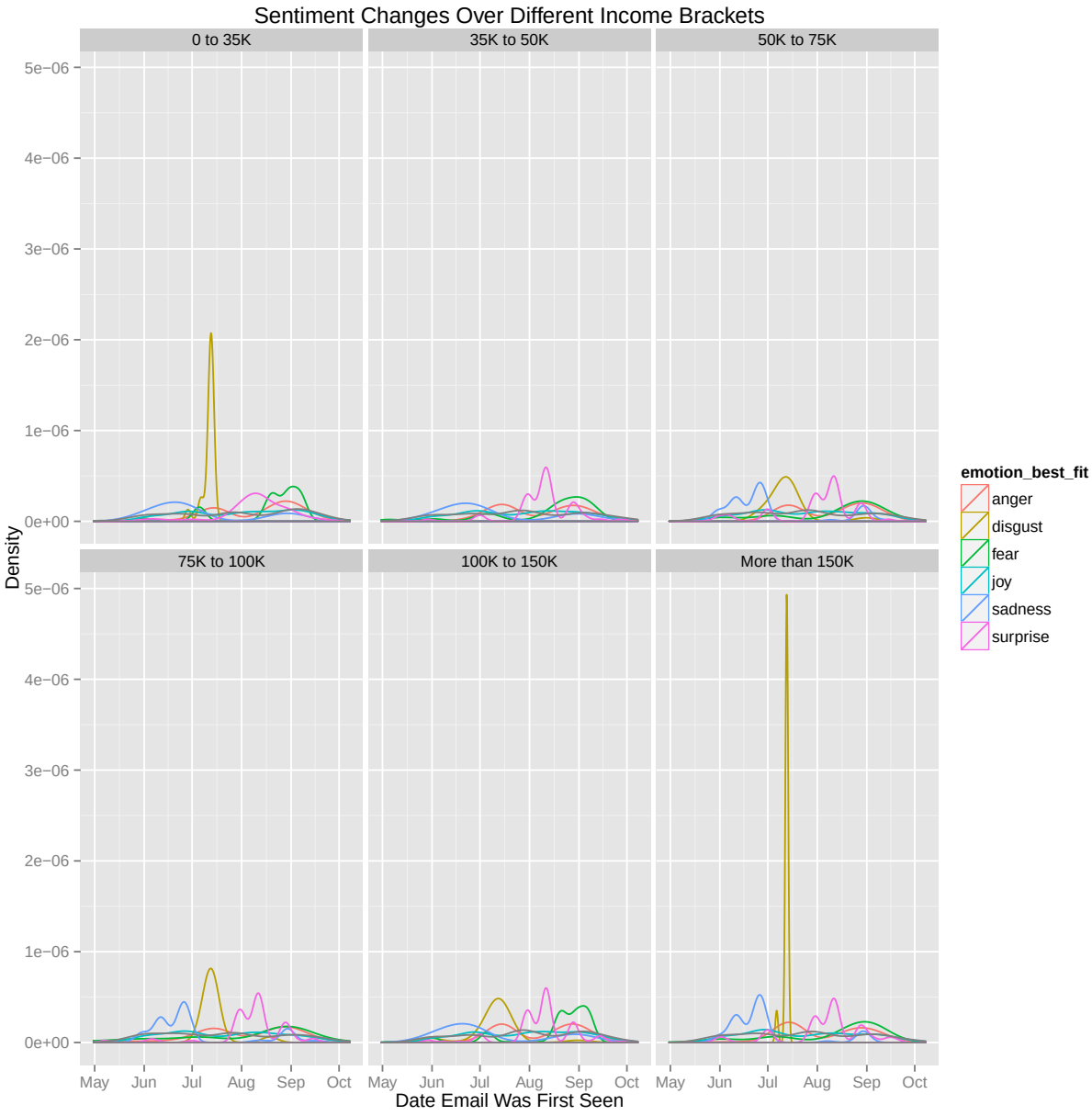

Figure 4: Faceted density plots showing the changes in sentiments regarding tax over time and for different income brackets, most noticeably the sharp rise in disgust in the ”0 to 35K” and ”More than 150K” income brackets. Note: Lines of different colors represent different sentiments.

significant events in the election. After seeing the above analysis and understanding that emails functioned as a tool to make people aware of Obama's policy and formally ask people to donate to the campaign. Twitter helped make President Obama, who was previously represented only through campaign emails and news, more personable and relatable.

Though we made some interesting discoveries, there are a few ways in which we could meaningfully expand on our sentiment analysis and overall findings. The first thing we could explore is alternative methodologies for parsing out sentiments from the email text. As stated previously in the methodology section, sentiment and natural language analysis is a technical field with its own set of complexities and subtleties that go beyond the scope of the `sentiment` package. In the future, we could experiment with alternative algorithms and conduct more research on the subject in order to derive more accurate sentiment metrics.

Our exploration would also benefit from a simply bigger analysis. For our first two findings, we could create a more interesting and relevant conclusion by merely just exploring more political events. This would provide information regarding both the reactions in the emails of the Obama for America campaign to each political event and further inductive evidence for our more general conclusions. It could also be beneficial to explore a larger data set, possibly one were all political parties were equally well-represented as the Obama campaign. In practice, this could involve finding a Republican-leaning email data set and adding it to the original.

Although we did our best to tie together the email and user data using `email_user`, a deeper and more wholistic analysis could be conducted if we had data concerning more individuals from parties outside of the Democratic tradition. This would allow us to see how opinions and sentiments about certain events varied across different parties based on their values. Having a larger group of people would likely also get rid of the shortage of emails discussing foreign policy, healthcare, or education.

In addition, we could expand on our Twitter analysis by not only looking at more information found on Twitter, such as how people re-post or re-tweet others' messages, but also including other written political communication mediums used by the Obama campaign, including other social networks. Then we could utilize the same sentiment analysis across even more mediums, allowing us to make broder comparisions.

We could also gather more information on foreign investment, healthcare, and education. Using just these words, "investment," "healthcare," and "education" was not enough to make conclusive arguments regarding the reality of complex political issues. Using additional key words or just finding additional data that also talks about foreign investment, healthcare, or education would provide us with much more data on these topics so we can consider the `registered_parties` unbiased and get a better idea of how sentiments change over time.

A Code

```
###
### Project 2
### Obama for America 2012: A Sentiment Analysis
###

# Load libraries for use in the project
library(ggplot2)
library(plyr)
library(stringr)
library(lubridate)
library(sentiment)
library(xtable)
library(twitteR)

###
### Load & Clean "Message Machine" email data set
###

# readRDS automatically loads strings as factors
email <- readRDS("email.RDS")

# Load the other data sets for later use
options(stringsAsFactors = FALSE)
user <- read.csv("user.csv")
email_user <- read.csv("email-user.csv")

# Join the user data with the email data
user_new <- join(user, email_user, type = "inner")
email <- join(email, email_user, type = "inner")

# Subset for Obama for America campaign
email <- subset(email, campaign == "Obama for America")

# Parse first_seen as a date with lubridate
email$first_seen <- ymd_hms(email$first_seen)

# Floor first_seen so that we can look over days
email$days <- floor_date(email$first_seen, "day")

# Some other useful time variables
email$wday <- wday(email$first_seen)
email$hour <- hour(email$first_seen)
email$month <- month(email$first_seen)
email$day <- day(email$first_seen)

# Most emails (>10) are not seen until 2012/04/30, so create subset
```

```
email <- subset(email, first_seen > ymd("2012-04-30"))

# Useful to omit NA's when looking at emotional best fit
email_na_omit_for_emotional <- subset(email, !is.na(emotion_best_fit))

# Create variables based on the sentiment and polarity of the emails
email$emotion <- classify_emotion(email$clean_text)
email$polarity <- classify_polarity(email$clean_text)

# Clean these new variables (polarity/emotion are lists)
email$anger <- as.numeric(email$emotion[ , 1])
email$disgust <- as.numeric(email$emotion[ , 2])
email$fear <- as.numeric(email$emotion[ , 3])
email$joy <- as.numeric(email$emotion[ , 4])
email$sadness <- as.numeric(email$emotion[ , 5])
email$surprise <- as.numeric(email$emotion[ , 6])
email$emotion_best_fit <- email$emotion[ , 7]

email$pos <- as.numeric(email$polarity[ , 1])
email$neg <- as.numeric(email$polarity[ , 2])
email$ratio_pos_neg <- as.numeric(email$polarity[ , 3])
email$polarity_best_fit <- email$polarity[ , 4]

# Clean repeated variables
email$emotion.ANGER <- NULL
email$emotion.DISGUST <- NULL
email$emotion.FEAR <- NULL
email$emotion.SADNESS <- NULL
email$emotion.JOY <- NULL
email$emotion.SURPRISE <- NULL
email$emotion.BEST_FIT <- NULL

email$polarity.NEG <- NULL
email$polarity.POS <- NULL
email$polarity.POS.NEG <- NULL
email$polarity.BEST_FIT <- NULL

email$polarity <- NULL
email$emotion <- NULL

email$X <- NULL
email$X.1 <- NULL

# Save this new data set as a csv file
write.csv(email, "email_clean.csv")
write.csv(email_na_omit_for_emotional, "email_na_omit_for_emotional.csv")

###
```

```
### Question 1 & 2: Sentiment Analysis over Time
###

# Plot email distribution over time for each sentiment
qplot(first_seen, data = email)
qplot(first_seen, data = email, fill = polarity_best_fit)
qplot(first_seen, data = email_na_omit, fill = emotion_best_fit)

# Political Dates of Interest
date1 <- as.numeric(ymd("2012-06-05"))  # Primaries in 6 states
date2 <- as.numeric(ymd("2012-06-28"))  # Supreme Court Ruling on Healthcare
date3 <- as.numeric(ymd("2012-07-28"))  # FEC Fundraising Deadline for Fall
date4 <- as.numeric(ymd("2012-08-28"))  # Republican National Convention
date5 <- as.numeric(ymd("2012-09-06"))  # Democratic National Convention
date6 <- as.numeric(ymd("2012-10-03"))  # First Presidential debate
dates <- c(date1, date2, date3, date4, date5, date6)
xtable(as.matrix(dates))

# Create timeline of important dates
timeline <- geom_vline(xintercept = dates, alpha = 0.5, size = 0.75)

# Create labels for timeline [numbers for first plot, letters for second]
label1 <- geom_text(data = NULL, x = date1, y = 1.59e-07,
  label = "1", size = 8, color = "red")
label2 <- geom_text(data = NULL, x = date2, y = 1.59e-07,
  label = "2", size = 8, color = "red")
label3 <- geom_text(data = NULL, x = date3, y = 1.59e-07,
  label = "3", size = 8, color = "red")
label4 <- geom_text(data = NULL, x = date4, y = 1.59e-07,
  label = "4", size = 8, color = "red")
label5 <- geom_text(data = NULL, x = date5, y = 1.59e-07,
  label = "5", size = 8, color = "red")
label6 <- geom_text(data = NULL, x = date6, y = 1.59e-07,
  label = "6", size = 8, color = "red")

labela <- geom_text(data = NULL, x = date1, y = 1.69e-07,
  label = "1", size = 8, color = "red")
labelb <- geom_text(data = NULL, x = date2, y = 1.69e-07,
  label = "2", size = 8, color = "red")
labelc <- geom_text(data = NULL, x = date3, y = 1.69e-07,
  label = "3", size = 8, color = "red")
labeld <- geom_text(data = NULL, x = date4, y = 1.69e-07,
  label = "4", size = 8, color = "red")
labele <- geom_text(data = NULL, x = date5, y = 1.69e-07,
  label = "5", size = 8, color = "red")
labelf <- geom_text(data = NULL, x = date6, y = 1.69e-07,
  label = "6", size = 8, color = "red")
```

```
# Density of polarity/emotion over time
qplot(first_seen, color = polarity_best_fit, data = email,
  geom = "blank", xlab = "Date", ylab = "Density",
  main = "Polarity Over Time") + geom_density(size = 1
  ) + timeline + label1 + label2 + label3 + label4 + label5 + label6
ggsave("project2-plot01-polariy-time.pdf",
  width = 11, height = 8.5, units = "in")

qplot(first_seen, color = emotion_best_fit, data = email_na_omit,
  geom = "blank", xlab = "Date", ylab = "Density",
  main = "Emotion Over Time") + geom_density(size = 1
  ) + timeline + labela + labelb + labelc + labeld + labele + labelf
ggsave("project2-plot02-emotion-time.pdf",
  width = 11, height = 8.5, units = "in")

###
### Question 3: Comparision of Twitter
###

# Load Tweets from Obama; convert list to data frame for ease of use
tweets <- userTimeline("BarackObama", n = 3200)
tweets <- twListToDF(tweets)

# Only able to recieve tweets starting in 2012-06-14,
# so subset emails to match the same date
email.twitter <- subset(email, first_seen > ymd("2012-06-14"))

# 2659 emails versus 2179 tweets since 2012-06-14

# Perform sentiment analysis of tweets and sort into new variables
tweets$emotion <- classify_emotion(tweets$text)
tweets$polarity <- classify_polarity(tweets$text)

# Clean sentiment variables
tweets$anger <- as.numeric(tweets$emotion[ , 1])
tweets$disgust <- as.numeric(tweets$emotion[ , 2])
tweets$fear <- as.numeric(tweets$emotion[ , 3])
tweets$joy <- as.numeric(tweets$emotion[ , 4])
tweets$sadness <- as.numeric(tweets$emotion[ , 5])
tweets$surprise <- as.numeric(tweets$emotion[ , 6])
tweets$emotion_best_fit <- tweets$emotion[ , 7]

tweets$pos <- as.numeric(tweets$polarity[ , 1])
tweets$neg <- as.numeric(tweets$polarity[ , 2])
tweets$ratio_pos_neg <- as.numeric(tweets$polarity[ , 3])
tweets$polarity_best_fit <- tweets$polarity[ , 4]

# Clean repeated variables
```

```
tweets$polarity <- NULL
tweets$emotion <- NULL

# Save tweets as a data set
write.csv(tweets, "tweets_clean.csv")

# Parse date and month as new variables
tweet$date <- ymd_hms(tweet$created)
tweet$month <- month(tweet$date)

# Density plots for emotion_best_fit
qplot(month, color = emotion_best_fit,
  data = email.twitter[!is.na(email.twitter$emotion_best_fit), ],
  geom = "density", xlab = "Month", ylab = "Density",
  main = "Sentiments of Obama For America"
  ) + xlim(4, 10)
ggsave(file = "project2-plot03-email.pdf", height = 6, width = 6)

qplot(month, color = emotion_best_fit,
  data = tweets[!is.na(tweets$emotion_best_fit), ],
  geom = "density", xlab = "Month", ylab = "Density",
  main = "Sentiments of President Obama's Twitter"
  ) + xlim(4, 10)
ggsave(file = "project2-plot04-tweet.pdf", height = 6, width = 6)

###
### Question 4: Case Study: Patient Protection and Affordable Care Act
###

# Search for emails related to tax or tax policy
email$tax <- str_extract(email$clean_text, "tax")
tax <- subset(email, email$tax == "tax")

# Which sentiments are most popular?
qplot(emotion_best_fit, color = polarity_best_fit, data = tax,
  geom = "histogram", xlab = "Best Fit Emotion", ylab = "Count",
  main = "Sentiments of the Patient Protection and Affordable Care Act")
# Joy is highest
ggsave(file = "project2-plot05-ppaca_hist.pdf", width = 6, height = 6)

# Order household income to rearrange facetted graphs in the future
tax$household_income <- ordered(tax$household_income,
  levels = c("0 to 35K", "35K to 50K",
  "50K to 75K", "75K to 100K",
  "100K to 150K", "More than 150K"))

# Plot changes in sentiments over time
qplot(first_seen, color = emotion_best_fit, data = tax, geom = "density",
```

```
  xlab = "Date Email Was First Seen", ylab = "Density",
  main = "Emotion Changes in Emails Regarding Tax")
ggsave(file = "project2-plot06-ppaca_emotion.pdf", width = 6, height = 6)

# Plot changes in sentiments over time for different income levels
qplot(first_seen, color = emotion_best_fit,
  data = tax[!is.na(tax$household_income), ], geom = "density",
  xlab = "Date Email Was First Seen", ylab = "Density",
  main = "Sentiment Changes Over Different Income Brackets"
  ) + facet_wrap(~ household_income)
ggsave(file = "project2-plot07-ppaca_income.pdf", width = 10, height = 10)

# Observations about plots:
# Spike in disgust in early July
# Probably had to do with people thinking that the Affordable Care Act
# mandate was going to translate into a tax
```