Correlation between "We the people" and Congress

Olivia Roy
University of Notre Dame
oroy@nd.edu

Timothy Blazek
University of Notre Dame
tblazek@nd.edu

Sung Hyun Shin
University of Notre Dame
sshin1@nd.edu

ABSTRACT

This paper deals with the concern of people feeling misrepresented by their respective Congress members. It is often hard for these members of Congress to know what the people they represent value; many insights come from phone calls and letters sent to them. However, in the modern day many people rely on social media to express their opinions rather than directly communicating with their representatives. It has also been debated if Congress is truly the branch of the people, since about 75% of the time, Congress members seem to simply vote on party lines [3]. Therefore, this paper will examine the relationship between how Congress members vote on certain issues and how the people they represent express their political views on these same topics via Twitter. Additionally, it will investigate if any Congress member's opinion change is in correlation with the people's current values. The challenges of this paper include location analysis and natural language processing. The results of this study will highlight potential strengths and weaknesses in the current legislative representation.

1 INTRODUCTION

The United States government is divided by three branches, the legislative, executive, and judicial branch. It is said that the legislative branch, composed of the Senate and the House of Representatives, is the branch of the people. When crucial and divisive votes are about to occur, people are often encouraged to call their representatives, or send them a letter. However, in this digital age, many people take to Twitter to express their political opinions. Congress members express their views on Twitter as well, which can often gain national attention. Because of the large use of Twitter in a political setting, the paper seeks to find if there is any correlation between the political views of US citizens on Twitter and how their respective Congress member votes.

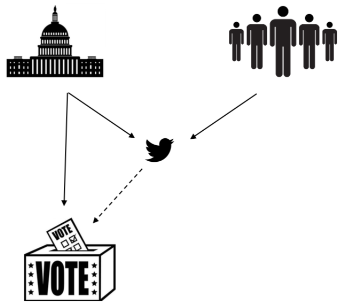

Figure 1: We know relationships between people and Twitter, Congress and Twitter, and Congress and votes exist. The paper attempts to discover the relationship between Twitter use and votes, as represented by the dotted line.

We are focusing on the July 25, 2017 Senate vote to repeal The Patient Protection and Affordable Care Act, also known as Obamacare. The vote failed 43-57, with nine Republicans siding with the Democrats. Because most Congress members vote with party lines, this vote is interesting because of these nine exceptions. One thing to note is that this vote was just for the repeal of Obamacare and not a replacement of Obamacare, so Republican members would be more inclined to not repeal Obamacare without a proper transition in place.

2 RELATED WORKS

Many researchers have used Twitter to measure public opinion on policy and politicians. One work compares measures of public opinion derived from polls to sentiments measured from tweet [11]. Another focused on the network structure of political microbloggers concerning their political parties and also political topics that they discuss [7].

Others are used to see how well public opinion reflects reality in elections. One analyzed voting intentions of users and the subsequent results in the French presidential and legislative elections of 2012 [5]. Another explored the relationship between tweet follows and shares gained by parties and the Delhi Assembly elections in India in 2015[13].

Different researchers looked at the Twitters of politicians and the relationship of characteristics of their Twitters and their received votes. One paper did a study where participants viewed a politician's Twitter page and found higher interactivity lead to more intentions to vote for them [10]. A different paper did content analysis of tweets by a politician and compared it to votes for the politician in the Dutch national elections of 2010, and found using Twitter in an interactive way resulted in more votes [9].

Our work is a new angle of research as it attempts to explore the relationship of Congressional voting and public opinion measured through sentiment analysis of tweets from these microbloggers.

3 METHODOLOGY

3.1 Data Collection and Cleaning

After finding which vote we would be looking at, the first requirement of the project is obtaining the Twitter data that is associated with the topic of the vote as well as picking an appropriate time frame before the vote was put on the floor. Using GetOldTweets [8], tweets containing the key word "Obamacare" were collected between July 19, 2017 and July 25, 2017. Additionally, the tweets both from and directed to the nine Senators who voted against their party were collected from July 1, 2017 to July 25, 2017. Next, we collected tweets from and directed towards five random Republicans and five random Democrats who voted along their party lines, again between July 1, 2017 and July 25, 2017. These ten Senators' tweets will be used as a control group. After these initial results, we collected tweets for every Senator to draw more conclusive results. For all of the Senators' tweets, they had to be further filtered by specific words such as "Obamacare" or "healthcare".

After collecting our tweets, we deleted the non-English tweets. Using langdetect [6], all of our tweets were passed through this language filter. The filter eliminated around 5.8% of the collected tweets with around 29.3% of those being false negatives. However, these false negatives tended to be only URLs without accompanied text.

3.2 Sentiment Classification

In order to get valuable information from the tweets, we had to figure out a way to quickly detect if the tweet was for or against Obamacare. To do this, we used sentiment analysis. We initially used the TextBlob library to analyze the tweets into two categories. Tweets with a score less than 0, are said to be negative, and correspond to voting for the repeal. Tweets with a score greater than 0, are positive and supportive of Obamacare, and thus be against the repeal. One challenge here is that two tweets can be for repealing Obamacare, but one is positive and one is negative in sentiment. For example, tweets like "Obamacare is death" and "We can do better than Obamacare" both support the repeal but have different sentiments.

Upon preliminary analysis as described below, we determined that additional work was needed on the sentiment analysis. We chose to train our own classifier into different group of tweets: positive, neutral, and negative categories. To do this, we first hand labeled 586 random tweets from the initial "Obamacare" tweets collected between July 19, 2017 and July 25, 2017. Of these tweets, 229 were positive, 87 were neutral, and 269 were negative. To keep about the same distribution, we randomly selected 170 positive tweets, 60 neutral tweets, 170 negative tweets to use as training data. We used the Python scikit-learn library to test Linear SVC, Multinomial Naive Bayes, Bernoulli Naive Bayes, and Complement Naive Bayes [12]. Additionally, we ran the TextBlob sentiment analysis on these labeled tweets to get a baseline result. Each of these five models were run 100 times and their accuracies were averaged. To use the tweet text as data for the models, we had to transform the data to a numerical representation. We did this using tf-idf, or term frequency-inverse document frequency.

One challenge with this project is that most tweets lack geolocation information. This makes mapping specific tweets and users to their correct representative difficult. One relationship we would like to analyze is the one between a Senator and the people they represent. To overcome this challenge, we will be focusing on tweets directed to Senators via the @ symbol. This is logical because these are most likely to be the tweets a Senator will see, and that could possibly influence the Senator.

The next step we took was to use the Complement Naive Bayes model, trained with the 586 labeled tweets, to classify the tweets directed towards Senators. We first used the classifier on the tweets for the nine Republican Senators who voted against the repeal. Additionally, we repeated this process with a control group of Senators. This control group was made of 5 randomly selected Democrats and 5 randomly selected Republicans who voted in favor of repealing Obamacare. This created a baseline for us to compare the results of the nine Republicans who voted against the repeal. Upon completing this comparison, we expanded our analysis to include all of the Senators. We took the average classified sentiment for each group: Republicans who voted Nay, Republicans who voted Yea, and Democrats and Independents who all voted Yea.

3.3 Justifying the Population

Because part of our question is if each Senator's vote is representative of their respective states, we needed to make sure the tweets we collected were also representative of each state. Thus, to validate our findings, we compared the number of tweets to a given Senator to the number of citizens in the corresponding state, which we got from US Census Data [1]. We did this by calculating $\frac{\text{number of tweets for the Senator}}{\text{population of the state}}$. We would expect there be more tweets from states with larger populations. The first step we took in calculating this was finding the population of each state and graphing it. This can be seen in Figure 2. The key states we want to note are California, Texas, Florida, and New York as the states with the most residents. We would also like to note that states Arizona, Colorado, Kentucky, and Alaska all have much smaller populations.

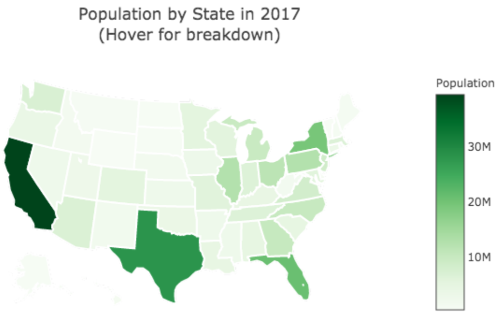

Figure 2: The shading of this map shows the population of each state in relation to all other states. The darker the shade of green, the more populous that state is.

Next, we graphed the number of tweets each Senator received and mapped them to their respective states. This can be seen in Figure 3. The most noticeable difference is how many more tweets the state of Kentucky received, when in comparison it Figure 2 it had a smaller population. We figured out this is because the two Kentucky Senators in 2017 were Rand Paul and Mitch McConnell, who was the Senate Majority Leader at the time. Both Senators are very popular, and thus received a lot of tweets. Additionally, Rand Paul is one of the Republicans who voted against the repeal of Obamacare. Overall, it is clear that these Senators skewed the number of tweets per population data.

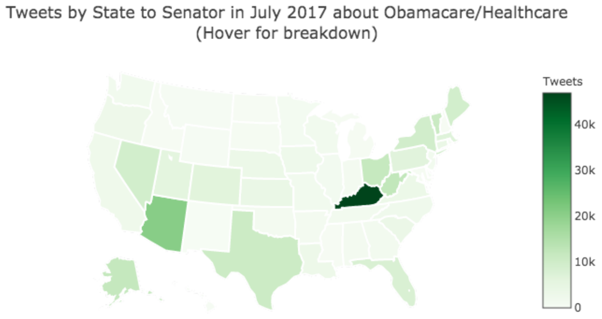
Tweets by State to Senator in July 2017 about Obamacare/Healthcare
(Hover for breakdown)
Tweets
40k
30k
20k
10k
0

Figure 3: The shading of this map shows the number of tweets of each state. The darker the shade of green, the more tweets the Senators of that state received.

When we were comparing the Republicans who voted Nay with all the other Senators, we noticed that they received more Tweets. The large amount of tweets Rand Paul and Mitch McConnell received inspired us to look at how the number of tweets that the nine Republicans received impacted our analysis of how many tweets come from each state. As seen in Figure 4, the Republicans who voted against party lines received significantly more tweets than the Senators that voted along their party lines. As you can see, 7 in 2000 people in the states where the nine Republicans were from, were likely to tweet at the Senator. Conversely, only 1 in 2000 people in every other state was likely to tweet at their Senator.

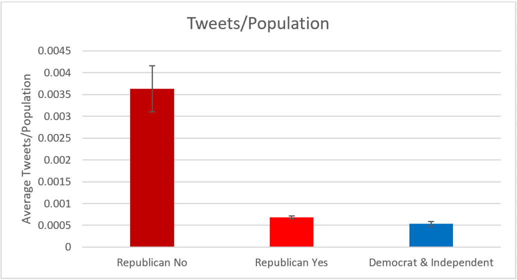

Figure 4: This graph shows the average number of Tweets divided by the population of each Senator's respective state. The Senators were grouped into one of the three groups: Republicans who voted No, Republicans who voted Yes, and Democrats and Independents.

This finding led us to believe that the overall distribution will be far similar to actual state populations without the tweets from the Republicans who voted Nay. Thus, we get the map shown in Figure 5. Here we can see that Texas, Florida, and New York have a higher number of tweets and they also have a large population. The two outliers are Arizona and Alaska. We think this is because John McCain was a Senator in Arizona and is well known throughout the country. The tweets in Alaska can be contributed to the state having a polarized view of healthcare because the two Senators, Lisa Murkowski and Dan Sullivan, had very different views of what it should look like.

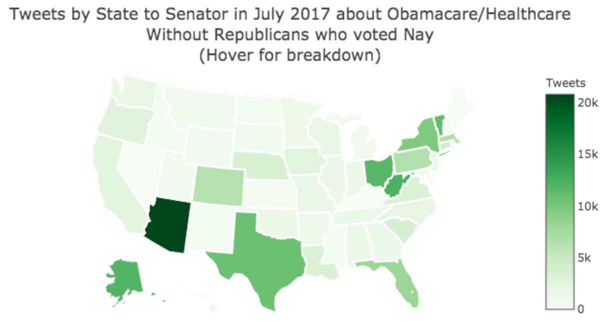
Tweets by State to Senator in July 2017 about Obamacare/Healthcare
Without Republicans who voted Nay
(Hover for breakdown)
Tweets
20k
15k
10k
5k
0

Figure 5: The shading indicates the number of tweets towards Senators of that State summed together. This map does not include tweets towards Republicans who voted Nay in repealing Obamacare.

Although the number of tweets per Senator mapped to their respective state does not match perfectly, it follows the population distribution to a degree of satisfaction. Namely, states with large populations, with the exception of California, received the most tweets, and states with smaller populations such as Hawaii, Maine, and states in the West Central Region, received fewer tweets. Thus, we feel our tweet data collection does represent the population of the United States.

4 EVALUATION

To test whether the Congress member's change their attitude according to the tweets they receive, we will attempt to classify their tweets first. After classifying the tweets, we will have a grasp of the sentiments of the tweets for each Senator. We can analyze these sentiments further to figure out whether the Congress member voted according to what the people were saying. We will classify the tweets by either pro-Obamacare, anti-Obamacare, or neutral towards the issue. We will evaluate the success of the classification by using the accuracy score, in other words, the percentage of tweets that it correctly predicted.

After we have an idea of the tweet classifications, we will want to compare them to Senators of similar parties or similar voting tendencies. We will want to isolate the nine Senators that voted against party lines so that we can get a better idea of how these nine are different from the other Senators. Afterwards, we will try to use this data to create a prediction model for how the Senators would vote. This prediction model can give us a better insight into the Twitter data that we collected because that model might pick up on Twitter tendencies that aren't as obvious. Furthermore, this prediction model can give us a better idea of whether the Twitter users can sway the vote or if the tweets give us a better idea of how the Senator is feeling about an issue.

5 ANALYSIS

5.1 Preliminary Analysis With Textblob

Using the GetOldTweets API, we gathered about 43,000 English tweets using the keyword "Obamacare" as described above. We then processed all of these tweets through the TextBlob and the results can be seen in Figure 6. Overall, we can see that that almost three-fourths of tweets were opposed to Obamacare, however, this number could be higher than it appears.

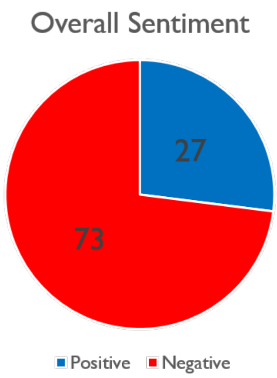

Figure 6: We can see from these initial sentiments that users are highly inclined to have negative tweets rather than positive.

One worry that we had was about certain Twitter accounts showing up too often in this query search. Therefore, we split up the results by username and counted how many tweets were negative and how many were positive for each user. From there, we counted how many users had a majority positive or negative attitude in their tweets. This brought us down to about 30,000 users with about a 76% of the tweets being negative about Obamacare. Surprisingly, the percentage increased despite eliminating duplicate users.

We then tried another tactic. Because we want to target how the people interact with politicians, we filtered out the texts that tagged Senators. We got a list of their handle tags from Twitter [2] and used these to track if people tagged any of them in the tweets. These tweets are more targeted from the people to bring about their opinion on issues so that it can influence the Senators' actions. This filtering was done alongside unique users and we got about 2586 users from this time frame that fit this. From these tweets, using TextBlob, we gathered that about 86% had a negative sentiment in their tweets.

Another thing that we tested was looking at who were being tagged and what they voted. We saw that there were 49 Republicans tagged and only 32 Democrats tagged. The sentiments score are shown in Figure 7. We can gather from this that when mentioning Senators, people tended to have more negative tweets towards Republicans than Democrats. Similarly, those who ended up voting in favor of Obamacare had 80% negativity, while those who voted against it had a 89% negativity rate. The sentiment here tells us that people were harsher towards the Republicans, who for the most part, voted against Obamacare. This initial method tells us that the vote swung in favor of the popular majority.

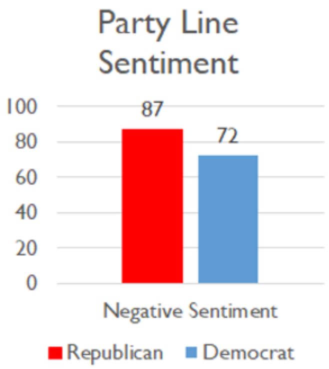

Figure 7: These results show us that people had harsher sentiments towards Republicans than Democrats by a 15% margin.

Overall, from the preliminary data, we notice that tweets about this political issue tend to be negative. We will do further analysis to determine whether their tweets were negative towards Obamacare or negative towards the politicians. Therefore, in order to investigate their sentiment towards Obamacare, we created a tweet classifier.

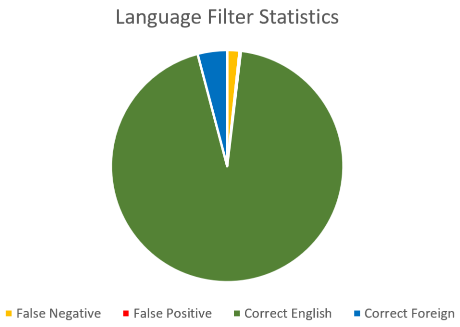

Figure 8: These show the statistics of the original 1000 tweets tested with the language filter.

Before we create a classifier, we had to remove non-English tweets to ensure an accurate classification and model. We performed a language filter on 1000 of these tweets to find statistics on how many tweets were deleted. We found that only 58 of those 1000 tweets were detected as foreign by langdetect. Of those 58, only 17 of those were false negatives, but most of these tweets were only URLs without accompanying text. We attempted to remove all links in tweets before detecting the language, but the overall numbers were negligible. In total, langdetect had a rate of 1.7% false negatives. Langdetect also failed to identify 2 foreign tweets, which gave it a rate of 0.2% false positives. As seen in Figure 8, false negatives weren't a relative issue and false positives are a minor issue. Once ran on the total 113,000 English tweets, we got a total of 108,000 English tweets, eliminating around 5% of tweets.

5.2 Tweet Classification Analysis

After classifying 586 tweets, which were randomly selected from the 108,000 tweets about Obamacare, we ran them through nine different models and calculated the accuracy, as shown in Table 1. From these model results, we gather that our previous concerns about using TextBlob's sentiment analysis as an inaccurate tool were correct and now we can use more accurate tools and models to classify people's tweets. We have two different metrics to determine which model behaves the best: accuracy and distribution. The accuracy metric determines the number of tweets guessed correctly, while the other metric looks at the distribution of what it predicted. We can see from this information that Complement Naive Bayes performs the best because it has the highest accuracy and it does not tend to favor one classification over another. Because of this, we will use Complement Naive Bayes to classify the rest of the tweets.

Table 1: Classification Testing Results

Model	Accuracy	Pro-Obamacare	Anti-Obamacare
Multinomial Naive Bayes	0.552	0.427	0.565
Bernoulii Naive Bayes	0.550	0.410	0.583
Complement Naive Bayes	0.577	0.430	0.490
Linear Support Vector Classification	0.549	0.388	0.542
Support Vector Machine	0.553	0.629	0.347
Decision Tree	0.463	0.434	0.413
Random Forest	0.524	0.565	0.367
Extra Trees Classifier	0.542	0.594	0.352
Text Blob Sentiment	0.265	0.311	0.541

Table 2 below shows the percentage of tweets directed towards a Senator that were positive and negative. These are the nine Republican Senators that voted against the repeal. We did also classify tweets in the neutral category, but they were minimal and thus not included in the table. The maximum percentage of neutral tweets any Senator received was 1.9%. There were just two Senators who received more support for Obamacare than negative feelings towards it: Dean Heller of Nevada and Lisa Murkowski of Alaska. However, these results are not too surprising. More people tend to speak out when they are passionate about a topic. Similarly, people tend to be more critical than supportive. Taking that into consideration, these results are about expected. These results will then be compared to that of the control group to draw more insightful conclusions.

Table 2: Classifying Tweets @Senators

Senator	Pro-Obamacare	Anti-Obamacare
Jerry Moran	34.1%	64.9%
Lindsey Graham	33.2%	66.5%
Rand Paul	28.3%	69.8%
Susan Collins	41.1%	58.1%
Bob Corker	37.5%	61.0%
Dean Heller	72.4%	27.2%
Mike Lee	23.2%	76.1%
Tom Cotton	38.5%	60.3%
Lisa Murkowski	59.9%	39.7%

Table 3 shows the data for the control group. The first thing we performed was a 2-sample t-test on the Democrats and the Republicans to see if there was a statistical difference between the two groups. The 95% confidence interval between the two suggest an interval of (0.03, 0.2). Because 0 is not in the interval, we can say that there is a statistical difference between the two. Because the Democrats had a higher anti-Obamacare average than the Republicans, we can say that the Democrats received higher proportions of anti-Obamacare tweets than the Republicans. We postulate that this is the case because people tend to say their mind if they disagree with something than if they agree. Therefore, the Democrats had more people tweeting at them with anti-Obamacare rhetoric.

Table 3: Classifying Tweets @Senators for the Control Group split by Republicans first and Democrats second

Senator	Pro-Obamacare	Anti-Obamacare
Roy Blunt	40.08%	59.23%
Richard Burr	46.3%	51.91%
Ben Sasse	32.86%	66.28%
Richard Shelby	42.59%	55.39%
Dan Sullivan	53.68%	45.48%
Bob Casey	29.98%	68.25%
Patrick Leahy	31.63%	68.37%
Bill Nelson	30.10%	68.69%
Gary Peters	30.64%	68.09%
Jon Tester	30.11%	62.5%

However, from this table we cannot say that there is a statistical difference from the Republicans who voted against party lines, whose data is detailed in Table 2 than from the Republicans who voted with party lines, detailed in the top half of Table 3. Even upon removing Murkowski as an outlier, the two still do not show a statistical difference at $\alpha = 0.05$ even though the average for them is higher than that of the control group Republicans. We cannot, therefore, say that these Republicans were swayed in one direction or another based off of the tweets directed towards them using this method.

Since our control group was made of five Republicans and five Democrats, we wondered if our findings could be applied across all of the Senators. Thus, we gave our classifier tweets directed towards each Senator and aggregated the results. Figure 9 below shows the average percentage of pro-Obamacare and anti-Obamacare tweets for Republicans who voted against the repeal, Republicans in favor of the repeal, and Democrats and Independents who were against the repeal. Indeed, we do see that the pattern we saw in the control group is present throughout all of the Senators. Firstly, Democrats, on average, have just under 70% of their tweets being classified as anti-Obamacare, and just under 30% of their tweets being classified as pro-Obamacare, leaving approximately 1.3% neutral tweets. Secondly, the difference between the two groups of Republicans, in respect to the ratio of pro-Obamacare tweets and anti-Obamacare tweets, is not significant. Thus, we can say that the ratio of tweets about pro-Obamacare or anti-Obamacare can tell us about what party Senators belong to, but we are limited and cannot say if they voted Nay or Yea simply based only on the sentiment of the tweets directed towards them.

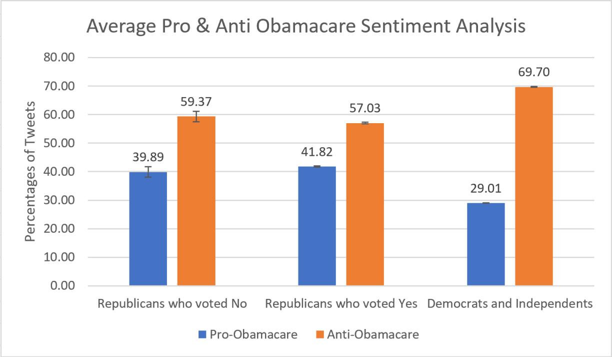

Figure 9: The average percentage of tweets classified as pro-Obamacare and anti-Obamacare. Neutrally classified tweets are not included. There is clear difference in the Democrat's and Independent's ratio of tweets compared to Republicans.

5.3 Vote Prediction

After getting the data for all 100 Senators, we wanted to create a prediction model using the tweet classification for each Senator and their Twitter traffic, which was normalized by their respective state population. Using different machine learning algorithms available through the sklearn package in Python, we got the data shown in Table 4. As you can tell from this table, the accuracy without feeding in the party's performs pretty well with a maximum accuracy of 76% using our custom ensemble method. This custom ensemble method takes in each of the other models' predictions and their accuracies. It weights its prediction based off of the confidence for each of the Senator's votes for each of the models and how accurate the model is to get a weighted democratic voting amongst all the classifiers. This helped average the different areas of weaknesses/strengths amongst the models to get a better, overall prediction model.

Table 4: Predicting vote using only Twitter Data

Model	Accuracy	Predicted Vote
MultinomialNB	0.738	31-69
BernoulliNB	0.565	0-100
LinearSVC	0.612	27-73
ComplementNB	0.745	33-67
Multi-Layer Perceptron	0.702	33-67
Decision Tree	0.686	43-57
Random Forest	0.704	38-62
Extra Trees	0.705	39-61
Custom Ensemble	**0.76**	**33-67**

However, this data from Table 4 only includes the gathered Twitter data, and does not include the party of the Senator, which is usually the best indication of how the Senator will vote. Table 5, shows the updated predictor, and we can see that the accuracy of the model increased for all the sklearn models applied as well as our custom ensemble. We can also see that the predicted vote was closer to the actual vote, which was 43-57. Overall, the predictor does a good job at predicting the final vote (it was only off by 2), but the accuracy is still only 88%. If we were to guess the vote solely off party lines we would obtain a 91% accuracy because only 9 Senators voted against party lines. Therefore, the accuracy would be higher if only looking at party lines, but the final predicted vote would be much worse than if we were to also use the Twitter sentiments and the tweet traffic for each Senator. Overall, we can say that the Twitter data adds more information about the voting tendencies of the politicians, but it doesn't give us the full story of how they will vote.

Table 5: Predicting vote using Twitter Data and Political Party

Model	Accuracy	Predicted Vote
MultinomialNB	0.780	34-66
BernoulliNB	0.876	54-46
LinearSVC	0.732	44-56
ComplementNB	0.793	39-61
Multi-Layer Perceptron	0.732	37-63
Decision Tree	0.844	52-48
Random Forest	0.825	48-52
Extra Trees	0.855	52-48
Custom Ensemble	**0.880**	**45-55**

Another piece of information we get from running the models in Table 5 are the Senators that it wrongly predicted. Using the Custom Ensemble method, we gather that it misclassified six of the people that voted against party lines as well as five Republicans and one Independent. However, this also means that it correctly predicted three of the people that switched their mind. We would like to get this number higher in the future, however, the overall prediction number gives us hope that the Twitter information gives us some insight into the voting.

6 LIMITATIONS

An initial limitation was how we classified tweets to train our model to identify pro and anti Obamacare sentiments. Our three members manually classified 200 different tweets each to train the model, and inconsistencies between our discretion could improperly train the model. This reduces our accuracy in identifying sentiments of tweets collected, and these sentiments were crucial in our project.

A clear limitation of this project was that our analysis focused on tweets towards the Senators due to the lack of geocodes. Without geocodes, we could not map the tweet to their individual Senator, which led us to analyzing the tweets at the Senators. This certainly left out some people's opinions regarding healthcare. A second limitation is that our model was trained on the set of tweets that contained Obamacare, not the tweets directed towards the Senators and that contained Obamacare. This was because we trained our model early in the project. Despite having different Twitter receivers, the sentiment decisions should be similar enough that the differences should be negligible.

A further limitation is that our search term of "Obamacare" and "healthcare" does not encapsulate all of the tweets we would want, such as people who tweet about the issue using similar words. Our search terms also included some tweets that we do not want, such as neutral tweets that could be tweets that just report news, tweets that do not really take a stance, or people tweeting about healthcare in other countries. Perhaps a better method of getting relevant tweets would include training a model to only select the relevant tweets.

Another limitation is that we only analyzed English tweets. This is undoubtedly excluding some citizens' point of view. We removed about 5% of our tweets due to our inability to train the model on non-English tweets. Similarly, some tweets that were removed by our language detector were links, photos, or videos. This too could have been valuable information that could have influenced our results. One other aspect we did not consider was bot or spam accounts. If they produced enough tweets, this too could have influenced our results in either direction. Finally, our model only had a prediction accuracy of 57.7%. Although this is significantly improved from TextBlob's predictions, it still has a large margin of error.

One of the biggest limitations was that we only looked at one vote. If we looked at more votes, we could have more basis for our conclusions and results. Our predictor would also perform better if a variety of votes could be analyzed. Then we would also have to create a more robust predictor that could predict different scenarios and votes.

7 CONCLUSION

In this paper, we inspected the relationship between Twitter and Congressional voting. We find no clear correlation between sentiment of tweets and the votes, but we did find patterns and trends that were utilized in our vote predictor that had achieved a 88% accuracy. If you predict every member votes with their party, you would get a 91% accuracy, but it would predict the wrong outcome of the vote. Our predictor, however, achieves a more correct outcome of a vote to not repeal Obamacare with a lower accuracy in respect to individual Senators. Thus, our predictor can be considered to perform better because it correctly predicts the outcome of the vote, while predicting with just party information would predict that the vote to repeal Obamacare would pass.

From our Figure 9, it would be logical to conclude that in the sense of representing their voters, Democrats and Independents aren't representing their people in their vote to not repeal Obamacare. That being said, only using the sentiment analysis of tweets are not completely indicative of public opinion, as there are many factors in why someone would tweet at a Senator more than others. For example, one factor would be that supporters would not be as vocal as dissenters because they realize that the Democratic and Independent members would not be willing to change their intended votes because it was the Democratic party that brought Obamacare into existence. Additionally, the Democratic members were not willing to repeal Obamacare because this vote did not include any program to replace it. Therefore, these factors make taking tweets at Senators an imperfect way to poll for overall public opinion.

In our predictor, the accuracy using just Twitter data from the tweets at the Senators was 76%. Once we feed in the information of what party the Senators were as well as the Twitter data, the accuracy jumps to 88%. This tells us that tweets do tell us a good amount of relevant information (enough to achieve 76% accuracy), but once outside data is also leveraged, we can more accurately predict the vote than using just outside data.

This paper aimed to find whether there was a correlation between the people of the United States and the Congress members that represent them. We can see from the tweets directed towards the Senators, that there is no direct correlation between the Twitter sentiments and the way the Senators voted. However, we found that the number of classified tweets have a correlation with the party that the Senators are associated with. Furthermore, because we created a predictor with acceptable accuracy, there are some patterns in the data that the predictor picked up regarding the voting tendencies and the Twitter sentiments.

8 FUTURE WORK

One of the first things that we would like to expand upon would be the votes that were considered. For this study, we only looked at one of the votes to repeal parts of Obamacare, however, there are many different votes that happen in Congress and in the House of Representatives every year in the United States. We would expand to include votes that Democrats switched their votes, Republicans and

Democrats switched their vote, and others that include the House of Representative voting. Another thing we could have looked at was the tweets further in the past for each of the politicians. For this study we looked at the 25 days leading up the vote, however, this was such a monumental policy that the politician's minds could have been swayed in the months leading up to this vote.

As for the model, we would have liked to look into the sentiment and the attitude of the Senator's Twitter handle shifted as it got closer to the vote. The best way to see how the politician will vote is to look directly at his vote and see if there is a shift in the mood over the course of time. Hence, our model could have handled these votes differently as to more accurately predict which way they will vote. Another way that we could train our model better is to look into the tweets themselves and handle more-liked or retweeted tweets differently. For example, we could have weighted the more liked tweets heavier in the model because those would be more likely to portray the popular opinion. Lastly, we could look into the links inside the tweets and analyze these more to get a better handle on the attitude of the user. For example, if the article heavily favors one side or the other, it would be simpler to classify the tweets using the attitude of the article along with than the user's commentary inside their tweet. Other things in the tweets that could be included are videos, images, and other links.

We could also added to the way that we obtained the data. For the purposes of this study, we only looked at tweets because Twitter has a good open source platform to obtain tweets. However, we could look at other social medias such as Reddit, Facebook, Instagram, or other medias of contacting a Senator such as hall towns, number of calls to the politician, or letters directed to them. This would enable us to get a better and fuller understanding of how a Senator's vote could be swayed by the public's opinion.

9 CODE BASE

All of our work is public in the GitHub repository found in the following link: https://github.com/oliviaroy20/SocialSensing

REFERENCES

[1] [n. d.]. Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2018 (NST-EST2018-01). ([n. d.]).

[2] [n. d.]. US Senator Twitter Accounts, 115th Congress. ([n. d.]). https://www.socialseer.com/resources/us-senator-twitter-accounts/

[3] 2018. American Government. (2018). http://www.ushistory.org/gov/6.asp

[4] 2018. Distribution of Twitter users in the United States as of September 2018, by age group. (2018). https://www.statista.com/statistics/192703/age-distribution-of-users-on-twitter-in-the-united-states/

[5] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16, 2 (2014), 340–358. https://doi.org/10.1177/1461444813480466

[6] Michal DanilÃąk. [n. d.]. ([n. d.]).

[7] Albert Feller, Matthias Kuhnert, Timm Sprenger, and Isabell Welpe. 2011. Divided They Tweet: The Network Structure of Political Microbloggers and Discussion Topics. (2011). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2759

[8] Jefferson Henrique. 2018. Get Old Tweets-Python. (2018). https://github.com/Jefferson-Henrique/GetOldTweets-python

[9] Sanne Kruikemeier. 2014. How political candidates use Twitter and the impact on votes. (2014). https://doi.org/10.1016/j.chb.2014.01.025

[10] Eun-Ju Lee and Soo Yun Shin. 2012. Are They Talking to Me? Cognitive and Affective Effects of Interactivity in Politicians' Twitter Communication. *Cyberpsychology, Behavior, and Social Networking* 15 (09 2012). Issue 10. https://doi.org/10.1089/cyber.2012.0228

[11] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. (2010). 11.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[13] Md Safiullah, Pramod Pathak, Saumya Singh, and Ankita Anshul. 2016. Social media in managing political advertising: A study of India. *Polish Journal of Management Studies* 13 (06 2016), 121–130. https://doi.org/10.17512/pjms.2016.13.2.12