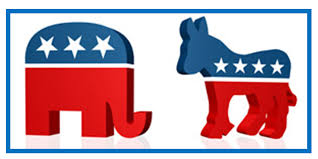
Prepared by: Ben Harwood, Chris Webster and Steve DeVito



Contents

[Introduction 3](#_Toc58579564)

[Analysis and Models 4](#_Toc58579565)

[About the Data 4](#_Toc58579566)

[Models 7](#_Toc58579567)

[Results 10](#_Toc58579568)

[Conclusions 14](#_Toc58579569)

[References 15](#_Toc58579570)

# Introduction

Imagine, for a moment, that you the reader were running for office. Obviously, it would be helpful to know how your potential constituents felt about you and your opponent so that you might have a good idea of your chance of winning. But how to do this reliably? You could certainly go old school and actually talk to people, but who’s got time for that? You have a campaign to run, after all!

Joking aside, gauging public sentiment toward political candidates is anything but easy. The tried and true method of polling seems to be losing its luster. On November 9, 2016, the Pew Research Center posted an article to its website who’s very first line was, “The results of Tuesday’s presidential election came as a surprise to nearly everyone who had been following the national and state election polling, which consistently projected Hillary Clinton as defeating Donald Trump.” (Deane, McGeeney, & Mercer, 2016) Leading up to the election that year, more than half of people told pollsters they believed Clinton would win compared to 27% who believe Trump would win, and while Clinton did win the popular vote that isn’t enough in our democracy, and late deciders in swing states ultimately overwhelmingly voted pro-Trump. It was conjectured that “shy Trump voters,” i.e. people who were embarrassed to say they planned to vote for Trump, affected the polls, yet a 2017 study done (again) by the Pew Research Center showed that was most-likely not true. (Kurtzleben, 2017) And as the last 6 weeks (as of December 10, 2020) have proven, the polls got things wrong again in the 2020. The media certainly hasn’t let this go unnoticed, with the left scratching their heads and the right essentially chastising them. While the goal of this study is not to come up with a real alternative to polling, all of this begs the question: if polling is no longer the de facto solution, where does one turn? Enter social media.

Volumes could probably be written on the affect social media platforms can and do have on elections. Indeed, Syracuse professor Dr. Jennifer Stromer-Galley has written extensively about this. (Enslin, 2020) But with the profound impact of social media, it is not unreasonable to think that perhaps it could be used to get a feeling for how people truly think about candidates, parties, and issues. It is an unfortunate truth that people lie, constantly, and even with this inherit truth, certain groups of people use the same language, talk about the same things, and almost seem to be of one mind in their tone, messaging, and beliefs. It is also possible that how they lie could betray them, in a sense that their true feelings lay under the surface. And there is possibly no better resource for seeing what people think about something than Twitter.

Even with a 280-character limit, a single tweet seems to almost be able to change the world. Ok that may be a stretch, but a tweet can certainly rile people up. The language of such a tweet, or even a less divisive one, has the potential to be quite telling about the individual who made it. It is not unrealistic to conjecture that the language in a tweet could be used to predict the political affiliation of the tweet[[1]](#footnote-1), and that a collection of tweets could be used to predict the affiliation of others. If this could be done reliably, how accurate are those predictions? Can such predictions be used to target people for donation requests and/or for exposure to topics and information that supports one candidate’s ideology over another? If voter ideology and strength can be predicted, perhaps it can be manipulated, especially for those that may be Independents or on the fence and open to other points of view.

This study seeks to answer some of those question.

# Analysis and Models

Before diving in, the general idea was to create a collection of tweets that could be easily identified as democrat or republican, use that data with various machine learning algorithms to see how well the party affiliation could be predicted just with the words use in the tweets, and then use the most effective model to predict the political ideological leanings of the write of tweets containing certain hashtags that have inherit polarity. This section will discuss the data collection process and the models used.

## About the Data

To begin, a dataset containing the tweets of the selected twitter accounts between November 1, 2020 and November 2, 2020 using the Twitter API was created. The twitter messages were auto labeled as Democrat or Republican based on each person’s/organization’s affiliation or obvious bias. An equal number of tweets was pulled from the ideological left and right by identify specific Twitter accounts for users that were representative of both sides. Included are politicians like Barack Obama and Mitt Romney; news channels such as MSNBC and Fox News; and pundits such as Rachel Maddow and Sean Hannity. There were 9 users from each side, with 250 tweets per user for a total of 4,500 tweets. This resulted in a collection of 11,389 tokens (after removal of stop words).

|  |  |  |  |
| --- | --- | --- | --- |
| Account | Label | Account | Label |
| housegop | Republican | SpeakerPelosi | Democrat |
| MittRomney | Republican | AOC | Democrat |
| seanhannity | Republican | donlemon | Democrat |
| lindseygrahamSC | Republican | SenWarren | Democrat |
| TuckerCarlson | Republican | Berniesanders | Democrat |
| tedcruz | Republican | barackobama | Democrat |
| foxnews | Republican | msnbc | Democrat |
| kellyannepolls | Republican | maddow | Democrat |
| realdonaldtrump | Republican | housedemocrats | Democrat |

Table 1 - Twitter accounts and labels assigned

Tweets are inherently messy, and users are limited by the number of characters they can use in a single tweet. Given this, as with most text mining projects, it was necessary to do some prepping, cleansing and reshaping of the data

Stop words were removed, using a custom list consisting of common lists from the Natural Language Toolkit (NLTK), sklearn, and other words specific to tweets, such as ‘rt’, ‘retweet’, and ‘amp’. A regular expression was also used to remove non-alphabetic words and word less than 4 characters in length. Finally, re-tweets and URLs were also removed. This was done with a custom-made function and was crucial for removing links.

The collected user tweets were then vectorized three ways, this allowed for extended testing to determine the best model. Specifically, they were vectorized using raw frequency counts, binary counts (i.e. whether the word was present in the tweet or not), and TFIDF for normalized counts. The vectorization resulted in three Document Term Matrices (DTMs). Several columns were added to each one:

1. Vectorizer label
2. Search parameter
3. Party – populated with “No-Detail” as a placeholder

Finally, the user tweet DTMs were consolidated into one data frame that was ready for exploratory data analysis (EDA) and modeling.

A picture containing table

Description automatically generated

Table 2 - Snippet of the final data frame

The first step in the exploratory data analysis was to look at the distribution of the vectorized corpus.

Most frequent words by count Most frequent words by number of documents

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Top 20 Words by Vectorizer

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Word | Frequency | Word | # Documents | Word | TDIF Mean |
| this | 896 | this | 790 | this | 0.026268 |
| that | 807 | that | 697 | that | 0.023133 |
| with | 596 | with | 556 | with | 0.020769 |
| have | 571 | have | 506 | have | 0.018699 |
| trump | 422 | trump | 399 | trump | 0.015266 |
| from | 409 | from | 382 | from | 0.015135 |
| will | 403 | will | 364 | will | 0.014409 |
| people | 393 | people | 360 | people | 0.014355 |
| about | 352 | covid | 324 | about | 0.012859 |
| covid | 337 | about | 317 | covid | 0.012498 |
| their | 335 | their | 302 | your | 0.012400 |
| they | 312 | president | 266 | their | 0.011877 |
| your | 305 | more | 255 | president | 0.011706 |
| president | 283 | election | 253 | election | 0.011483 |
| more | 283 | they | 253 | more | 0.011327 |
| election | 273 | your | 252 | they | 0.010918 |
| vote | 262 | americans | 236 | vote | 0.010596 |
| americans | 250 | american | 233 | american | 0.009485 |
| american | 239 | vote | 221 | americans | 0.009442 |
| house | 219 | house | 212 | house | 0.009035 |

Regardless of the vectorizer used, the top 20 words were very similar. The next step was to create a variety models with the vectorized corpuses and compare them to see which ones should be further scrutinized to be candidates for the final model.

Roughly the same process was followed to create an unlabeled data frame consisting of 250 tweets for each of the following hashtags:

1. #trump2020
2. #maga
3. #biden2020
4. #bidenharris

This data frame was unlabeled so that the most effective model could be applied to make the party labels.

## Models

Numerous machine learning algorithms were employed to measure the predictability of the party labels, namely multinomial and Bernoulli Naïve Bayes, support vector machines, decision trees, and random forests. These were each applied with each of the three vectorizers to a randomized 80% of the data used for training. The following table shows the nine most effective models based on their performance on the remaining 20% of the data used for testing.

Model Train/Test Accuracy (one-fold)

Model Vectorizer Train % Test %

BernoulliNB CountVectorizer (Binary) 93.519 81.444

Multinomial NB CountVectorizer 95.593 81.278

Multinomial NB CountVectorizer (Binary) 95.778 80.889

-------------------------------------------------------------------------- cut-off

RandomForest CountVectorizer (Binary) 99.185 78.944

SVC (Linear) CountVectorizer 99.037 78.278

RandomForest CountVectorizer 99.259 77.889

RandomForest TF-IDF 99.222 77.722

SVC (Linear) TF-IDF 95.926 76.944

Multinomial NB TF-IDF 93.481 75.333

In general, the Binary Count Vectorizer produced the best accuracies with the Bernoulli Naïve Bayes algorithm at the top. The Multimomial Naïve Bayes algorithm was second best. Overall, the differences between the top three model accuracies was negligible with a range of 0.555. The Support Vector Machine algorithm was not as accurate as the others and the TFIDF Vectorizer was the least accurate compare to the CountVectorizers.

This below table shows the same data after four-fold cross-validation.

Cross-Validation Results (four-fold):

Train % Test %

Model Vectorizer (avg) (avg)

Multinomial NB CountVectorizer 95.61 81.99

Multinomial NB CountVectorizer (Binary) 95.82 81.13

BernoulliNB CountVectorizer (Binary) 93.34 81.01

-------------------------------------------------------------------------- cut-off

SVC (Linear) CountVectorizer 98.99 78.91

Multinomial NB TF-IDF 94.22 77.47

RandomForest CountVectorizer (Binary) 99.30 77.26

RandomForest CountVectorizer 99.32 77.19

RandomForest TF-IDF 99.24 76.91

SVC (Linear) TF-IDF 95.90 76.87

After cross-validation, the top three models were consistent, in terms of average test accuracy, with the initial runs and were selected for further analysis.

Learning Curves

A Learning Curve shows whether the model can benefit from more data or lower complexity. It shows the relationship of the training score versus the cross validated test score for an estimator with a varying number of training samples. This visualization is typically used to show two things:

1. How much the estimator benefits from more data (e.g. do we have "enough data" or will the estimator get better if used in an online fashion).
2. If the estimator is more sensitive to error due to variance vs. error due to bias.

If the training and cross-validation scores converge together as more data is added, then the model will probably not benefit from more data. If the training score is much greater than the validation score then the model probably requires more training examples in order to generalize more effectively.

The curves are plotted with the mean scores, however variability during cross-validation is shown with the shaded areas that represent a standard deviation above and below the mean for all cross-validations. If the model suffers from error due to bias, then there will likely be more variability around the training score curve. If the model suffers from error due to variance, then there will be more variability around the cross validated score.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

The Bernoulli Naïve Bayes model appears to have noticeable sensibility to error due to bias than the other models. All the models appear to be sensitive to error due to variance. The Multinomial Naïve Bayes model training and test curves appear to be converging so they probably would not benefit from more data. However, the Bernoulli Naïve Bayes model training and test curves are not clearly converging so more data and/or less complexity might help.

Visualize Receiver Operating Characteristic Curves

ROC curves are typically used in binary classification to study the output of a classifier and typically feature the true positive rate on the Y axis, and the false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

The “steepness” of the ROC curve is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

All the ROC curves were similar in a good way. There is a large AUC and the True Positive rate is clearly maximized over the True Negative rate.

Using the methods of Cross-Validation, Learning Curves and ROC/AUC, the Multinomial and Bernoulli models were comparable and warranted further analysis from a test/training perspective.

# Results

Confusion matrices and classification reports visualize the model results predicted from the testing data set compared to their actual classes. In addition, Accuracy, Precision, Recall and F1 metrics help with interpreting a model’s efficacy.

* Recall should be as high as possible. It indicates how many of the positive classes were predicted correctly.
* Precision indicates how many of the predicted positive classes are actually positive.
* Accuracy should also be as high as possible. It indicates how many predictions were correct out of all the classes.
* The F1-score uses the harmonic mean of Recall and Precision to make comparison between models simpler.

Multinomial NB – CountVectorizer

Accuracy: 81.278%

Classification Report for Multinomial NB - CountVectorizer:

=================================================================

| | Precision | Recall | F1-score | Support |

|-------------|-------------|----------|------------|-----------|

| Democrat | 0.7742 | 0.8687 | 0.8187 | 876.0000 |

| Republican | 0.8592 | 0.7597 | 0.8064 | 924.0000 |

=================================================================

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

True Positives (Democrats): 761

True Negatives (Republicans): 702

False Negative (Republicans): 222

False Positives (Democrats): 115

Multinomial NB – CountVectorizer (Binary)

Accuracy: 80.889%

Classification Report for Multinomial NB - CountVectorizer (Binary):

=================================================================

| | Precision | Recall | F1-score | Support |

|-------------|-------------|----------|------------|-----------|

| Democrat | 0.7694 | 0.8779 | 0.8201 | 893.0000 |

| Republican | 0.8604 | 0.7409 | 0.7962 | 907.0000 |

=================================================================

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

True Positives (Democrats): 784

True Negatives (Republicans): 672

False Negative (Republicans): 235

False Positives (Democrats): 109

BernoulliNB – CountVectorizer (Binary)

Accuracy: 81.444%

Classification Report for BernoulliNB - CountVectorizer (Binary):

==================================================================

| Precision | Recall | F1-score | Support |

|----------------------------|----------|------------|-----------|

| Democrat | 0.8584 | 0.7486 | 0.7998 | 891.0000 |

| Republican | 0.7810 | 0.8790 | 0.8271 | 909.0000 |

| avg / total | 0.8193 | 0.8144 | 0.8136 | 1800.0000 |

==================================================================

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

True Positives (Democrats): 667

True Negatives (Republicans): 799

False Negative (Republicans): 110

False Positives (Democrats): 224

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Important Features | | | | | | | | | | | |
| **Multinomial NB - Count Vectorizer** | | | | **Multinomial NB - CountVectorizer (Binary)** | | | | **BernoulliNB - CountVectorizer (Binary)** | | | |
| **Democrat tweets** | | **Republican tweets** | | **Democrat tweets** | | **Republican tweets** | | **Democrat tweets** | | **Republican tweets** | |
| this | -4.5317 | this | -5.0230 | this | -4.5673 | this | -5.0181 | this | -1.4523 | this | -2.0138 |
| that | -4.8148 | that | -5.0928 | that | -4.7008 | with | -5.0830 | that | -1.6645 | that | -2.1707 |
| with | -5.1624 | with | -5.1039 | have | -5.0464 | that | -5.1705 | have | -1.9548 | with | -2.1972 |
| have | -5.1891 | have | -5.3731 | with | -5.1892 | have | -5.3431 | with | -2.0357 | have | -2.3729 |
| people | -5.2737 | from | -5.5235 | trump | -5.3433 | from | -5.4664 | trump | -2.1612 | from | -2.5190 |
| trump | -5.4536 | will | -5.5852 | people | -5.3495 | trump | -5.5883 | people | -2.2758 | will | -2.7014 |
| your | -5.5042 | trump | -5.6318 | will | -5.5634 | will | -5.6957 | will | -2.3891 | trump | -2.7014 |
| about | -5.5419 | they | -5.7110 | about | -5.5711 | president | -5.7817 | from | -2.3891 | covid | -2.7823 |
| will | -5.6473 | about | -5.7859 | from | -5.5711 | they | -5.7930 | covid | -2.4134 | their | -2.8704 |
| from | -5.6473 | people | -5.9163 | your | -5.5711 | people | -5.8637 | their | -2.4383 | american | -2.9107 |
| vote | -5.6559 | president | -5.9163 | their | -5.5789 | american | -5.8759 | about | -2.4383 | about | -2.9245 |
| election | -5.6645 | their | -5.9419 | election | -5.5947 | more | -5.8884 | your | -2.5827 | democrats | -2.9245 |
| their | -5.6999 | biden | -5.9550 | covid | -5.6434 | covid | -5.9397 | americans | -2.5925 | president | -2.9385 |
| they | -5.7090 | more | -6.0091 | vote | -5.6688 | about | -5.9397 | election | -2.6427 | people | -2.9385 |
| covid | -5.7090 | democrats | -6.0372 | they | -5.6860 | democrats | -5.9663 | president | -2.6741 | more | -2.9527 |
| americans | -5.7273 | your | -6.0372 | americans | -5.6947 | biden | -5.9663 | vote | -2.6957 | biden | -2.9671 |
| more | -5.8043 | american | -6.0810 | health | -5.7675 | their | -5.9799 | more | -2.6957 | senate | -3.0580 |
| make | -5.8877 | covid | -6.0961 | more | -5.8258 | your | -6.0363 | they | -2.6957 | they | -3.0903 |
| health | -5.9321 | what | -6.1914 | make | -5.8983 | senate | -6.0807 | make | -2.7631 | house | -3.1068 |
| need | -5.9321 | election | -6.2253 | care | -5.9536 | over | -6.1273 | health | -2.7986 | election | -3.1407 |

Table 3- Most important features (log\_prob)

The above chart shows the most important features between the Democrat and Republican tweets in the entire corpus across the top three algorithm/vectorization combinations. The words highlighted in red are unique to the Democrats and Republicans within a model. This does show some uniqueness between the words used in tweets by known Democratic public figures versus their Republican counterparts.

It was a tough call but the Multinomial Naïve Bayes, using the pure frequency vectorizer, was the most accurate on the test set, so it was decided to use that one for the final hashtag analysis. Specifically, additional twitter data was pulled for hashtags that were representative of Democrat and Republican ideologies as mentioned in Section 1.1. There were 250 . The following word clouds represent some of those hashtags:

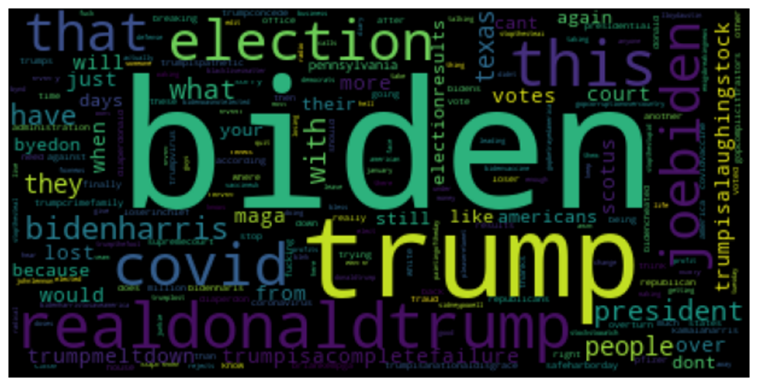
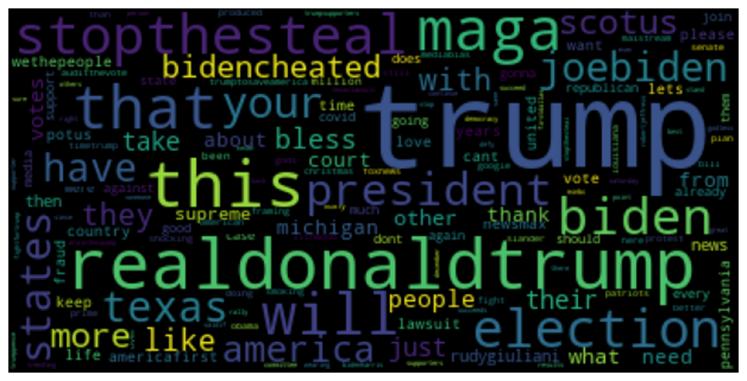


Figure 1 - #trump2020 Figure 2 - #biden2020

Text

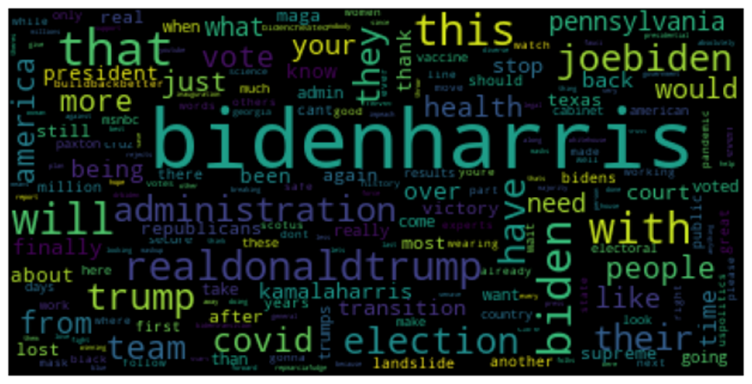
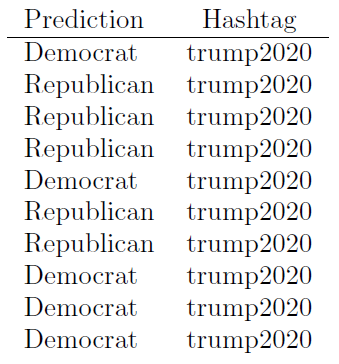
Description automatically generated

Figure 3 - #maga Figure 4 - #bidenharris

Some curiosities arose, for example, “trump” is the second most common word among the #biden2020 tweets. Overall, the final model predicted that 561 of the hashtag tweets were Republican and 439 were Democrat. The below table shows what was expected, that hashtags related to one side might be more reflective of the other party.

Going a bit deeper,

* the #biden2020 tweets were 146 Republican to 140 Democrat
* the #bidenharris tweets were 77 Rebublican to 173 Democrat
* the #maga tweets were 194 Republican to 56 Democrat
* the #trump2020 tweets were 144 Republican to 106 Democrat.

Again, this was not overly surprising, except that more of the #biden2020 tweets were labeled Republican than Democrat. And here’s a really interesting example:

“Trump lied to MAGA and the American people Demand a CovidCommission Coronoavirus updates Every hour Americans are dying AstraZeneca plans additional trial after error hospitalizations hit record again via usatoday”

While the above tweet has MAGA included, almost a conservative rallying, the tweet is not exactly a positive use of the slogan. The model predicted, the authors believe correctly, this to be a democrat tweet.

# Conclusions

Text classification can be a powerful and enlightening tool for extracting knowledge and wisdom from unstructured social media data such as Twitter tweets. Twitter is now culturally embedded and ubiquitous as an information source evidenced by its millions of users and sheer volume of tweets published on a daily basis.

It is now possible to use published tweets to approximate what a person’s political ideology is using widely accepted classification methods. With the tools available, thousands of tweets can be retrieved and analyzed in just a few minutes to categorize and classify them along ideological lines.

After this study, we found that peoples’ tweets can be used to classify them as Demcorat or Republican, and that the results were somewhat accurate. With improved sentiment analysis algorithms that can better detect sarcasm and emotion and more data cleaning/tokenization based on the language and patterns of tweets, we could use them to better predict the next presidential election results in 2024.

One challenge with tweets is that they can be posted by anyone or anything (e.g. bots) and can be constructed to be purposely misleading or contain alternative facts (e.g. fake news) which can confuse the results.

In 2016 and 2020, the pollsters leading up to the election got things very wrong. But was it their fault, or did people just “say the right thing”? Better yet, could what was explored here have prevented the botched polling predictions? With further analysis, tuning, and employing additional methods, we think yes!

# References

Chouhbi, Kamal (2020, February 5). *Are you Democrat or Republican? Let your tweets define you …* Retrieved from Towards Data Science: <https://towardsdatascience.com/are-you-democrat-or-republican-let-your-tweets-define-you-4aa4cadf4bea>

Deane, C., McGeeney, K., & Mercer, M. (2016, November 9). *Why 2016 election polls missed their mark.* Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/

Enslin, R. (2020, April 29). *Syracuse University.* Retrieved from Disinformation and the 2020 Election: https://www.syracuse.edu/stories/jennifer-stromer-galley-2020-presidential-election/

Kurtzleben, D. (2017, May 5). *Pollsters Find 'At Best Mixed Evidence' Comey Letter Swayed Election.* Retrieved from npr: https://www.npr.org/2017/05/05/526936636/pollsters-find-at-best-mixed-evidence-comey-letter-swayed-election

1. That’s actually what they’re called [↑](#footnote-ref-1)