Analyzing a Presidential Campaign

BIDEN

PRESIDENT

Michael Chen, Rebecca Driever, & Rayna Ji

* This is an exercise in performing analysis for a campaign, **not** a reflection of any individual's political views and **not** an endorsement of any candidate for President

Table of Contents

Executive Summary

In this project, we dived into the 2020 Democratic Primary election and focused on Joe Biden and his biggest competitor Bernie Sanders. We are committed to finding out the correlation between candidate's performance with media reports mentions and fundraising. To achieve that, we did topic modeling, sentiment analysis and generated word-clouds by using the unstructured data of the New York Times articles since Feb 1 until March 2. We employed logistic regression to find out the accuracy of predicting delegate-won results from fundraising information.

The topic modeling gave us an idea about what is mainly talked about regarding Joe Biden and Bernie Sanders.

The sentiment analysis was achieved by calculating scores of polarity and objectivity. We found that on average articles mentioning Joe Biden more than Bernie Sanders tended to be more negative (lower polarity score), while mean objectivity scores didn't differ and tended to be on the more objective side (less than 0.5).

Regarding the correlation between candidate's performance in winning a state with fundraising, the logistic regression was applied three times on Biden's dataset, Sanders' dataset and an aggregated data set which gave respective accuracies of 0.571, 0.778 and 0.667. This indicates that it can be useful for a candidate campaign team to predict whether the candidate can win a state by looking at how many donors were there in the state and how much they contributed. We also provided suggestions for further improvement on this analysis about collecting and considering more marketing data into the model, like TV ad spendings, number of rallies held, number of direct mail pieces sent out, etc.

Background and Context

On April 25, 2019, former Vice President Joe Biden announced his bid for the Democratic Party candidacy in the 2020 Presidential election[1]. Prior to the first ballots of the primaries being cast, Biden was thought to be a strong front-runner alongside Bernie Sanders[2]. However, disappointing finishes in Iowa and New Hampshire made the press question whether his campaign could make it to Super Tuesday -- a day when ⅓ of the total delegates are decided. As a former VP, Biden entered the race with a lot of name recognition. However, name recognition is not always the ticket to the nomination, as his former boss Barack Obama proved in the 2016 Presidential race.

Securing the nomination is said to be based on the three M’s: money, momentum and movement (organization)[3]. Most campaigns keep track of their money very well - where it’s coming from, who it’s coming from, and where it’s being spent. Movement is often determined by knowing one’s own candidate base and knowing where a campaign should set themselves up ahead of primaries in order to have the greatest impact on votes. Momentum is difficult to predict, but can be inferred by monitoring news outlets and public opinion.

Discussion of Traditional Approaches

In the candidates’ campaign teams, financial information is typically tracked in a structured fashion such as a database for easy analysis, however media coverage is not. Most likely, one individual is responsible for monitoring reporting sources such as the New York Times and NPR, while another is responsible for monitoring social media trends and opinions. These individuals will often need to go directly to the media outlets and manually search through a large amount of information, and use personal judgement regarding the content of these sources. With manual search and eyeballing the content can be challenging and extremely hard for the campaign staff. Most of the stories or details will likely be lost or skipped over. Traditional overler-addressed analysis on financials, polls, and rally activities statistics are due to ease of access to the data and less technical capabilities required. This often leads to a limited comprehension of how major news media, social media and campaign operation works together while impacting each other.

However, the approach proposed in this report allows easy monitoring of candidate mentions in the news. There are over one thousand New York Times articles written each week, especially during campaign season -- in the month of February there were over 700 articles that mentioned either Joe Biden or Bernie Sanders. Leveraging automated data pipeline and analysis on the New York Times news articles no longer need for a campaign staffer to physically search and sort through the thousands of articles written every month. Any analysis that can be performed on articles, delegates, or financial contributions regarding the Biden campaign can just as easily and efficiently be performed on the Sanders campaign. This can provide a great advantage to the Biden team if utilized.

Model Building and Analyses

The main analysis done by our team was Natural Language Processing on New York Times articles, as this is less likely to be done by campaigns currently. Topic modeling and sentiment analysis were performed on the text data, while a logistic regression was performed on financial information from the FEC and delegate counts of states that have already held their primaries.

Topic Modeling

Topic modeling was done using both information from the NYT API as well as the article text itself. Rather than going through each article looking at the entire body of text, we first wanted to determine how much information could be garnered from doing topic modeling on article headlines. After looking at headlines, topic modeling was done on the entire bodies of articles. In total, 839 articles from February 1st, 2020 through March 2nd, 2020 were analyzed, each mentioning either Biden or Sanders at least once.

Before any topic modeling can be done, text data must be preprocessed. This preprocessing technique was the same for both headlines and body text, consisting of the following steps:

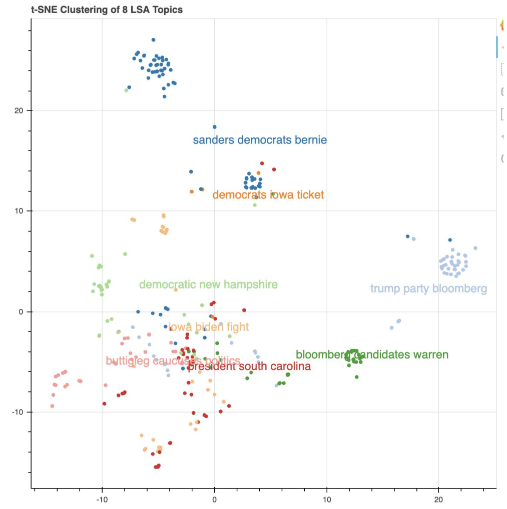
1. Convert all words to lowercase
2. Handle negation by changing all “xxxn’t” contractions to their root and “not”, then combine “not” with the word following it
3. Remove stop words and punctuation
4. Stem words using Porter’s 1979 stemmer for the article body text.

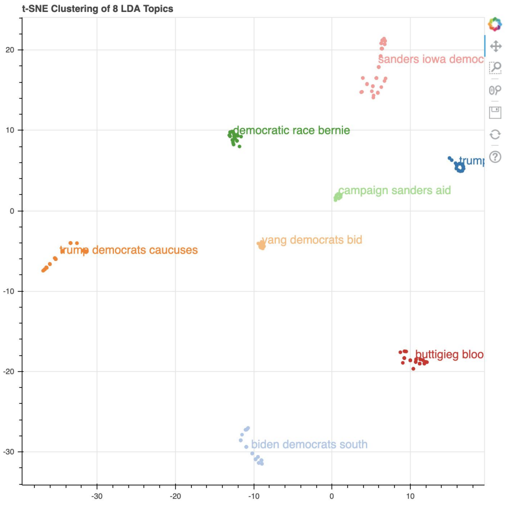
Topic Modeling for News Articles’ Headlines

A word cloud based on the frequency of the words in headlines is plotted at the beginning. From this very initial exploration (Appendix A), we can already see “Sanders” being heavily weighted in NYT’s headlines. We then proceed with topic modeling to gain more insights on how these headlines can be classified into different topics.

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are the two major techniques used in this analysis to understand text by modeling topics it represents. Both techniques are unsupervised learning, where headlines with similar “topics” will be clustered together. LSA works similar as PCA does by only keeping the most significant dimensions in transformed space. LDA works under Bayesian assumptions, where it looks at the probability distribution of certain words given the distribution of a particular topic. The main goal is to compare the two techniques and lead our way into the following sections, in which the magnitude and the complexity of the raw text increases dramatically.

In LSA (Appendix A), out of eight topics, two contained “biden” within the top 10 words. On the other hand, “sanders” appeared only once out of the eight topics. This however is counter intuitive from what we saw from the word cloud mentioned above. Meanwhile, in LDA clusters, out of eight topics, “sanders, bernie” showed up in four, whereas “biden” only showed up twice. LSA’s clustering suggests that NYT favors “biden” with underlying topics analysis, however LDA suggests the other way. We then plotted both clusters with 2D representations using t-SNE to assess its corresponding validity.

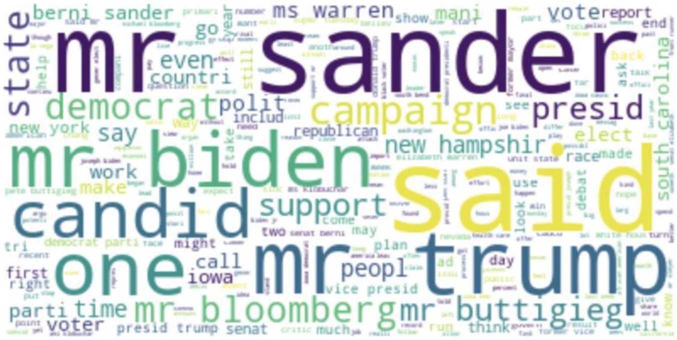




Based on the above graphs, we can conclude that LDA performed much better in terms of topics classification, and it also aligns with our expectation from the word cloud. Therefore, we will use LDA in the following analysis as well. From the same graph, Senator Sanders' name also appeared in headline topics alongside the other former Democratic candidates Andrew Yang and Elizabeth Warren, as well as President Trump, while for the topic where Biden was a top word Sanders was also prevalent. In addition, we notice that the cluster for "biden" is distant from other clusters. This may suggest that the Times writes differently about Joe Biden than the rest of the candidates.

Topic Modeling for Article Body Text

After seeing these trends in just the headlines of articles, our team decided to look deeper at the full body text to determine if there were any clearer ideas of topics that Vice President Biden was mentioned in, as well as see how the trend in prevalence of his versus Senator Sanders' name held throughout the text. Since this text requires more analysis, stemming was done which caused "sanders" to become "sander" and "bernie" to become "berni". An initial word cloud shows that while Senator Sander's name is clearly more prevalent than Vice President Biden's, their counts are much closer than when looking at headlines.



When topic modeling was done on the article body text, words were classified into 10 topics and the top 20 words in each topic were examined to get an idea of what each topic represented. Vice President Biden's name was seen in topics 0 and 4, while Senator Sanders' name was seen in topics 1, 4, 6, 8 and 9. Topic 0 mentioned words like "campaign", "senate", and "family" in addition to Mr. Biden, with nothing standing out as a clear idea of what this topic includes. Topic 4, on the other hand, mentioned not only Sanders and Biden, but also Buttigieg and Warren, along with New Hampshire and Iowa -- most likely about those two first primary states. Exact words in each topic can be found in Appendix B.

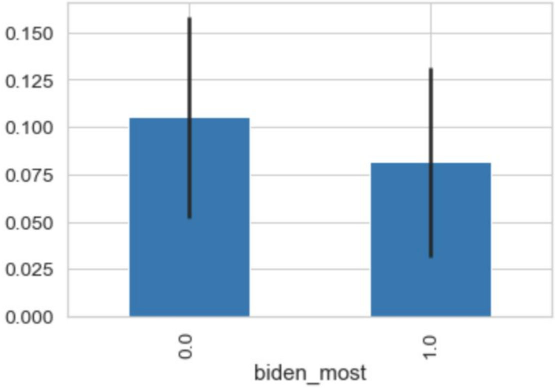
Sentiment Analysis

The New York Times, a leading source of news media for American Democrats, has not officially endorsed a candidate in the Democratic primaries. However, our team decided to perform sentiment analysis on the article body text to determine if the news source is indirectly showing support one way or another.

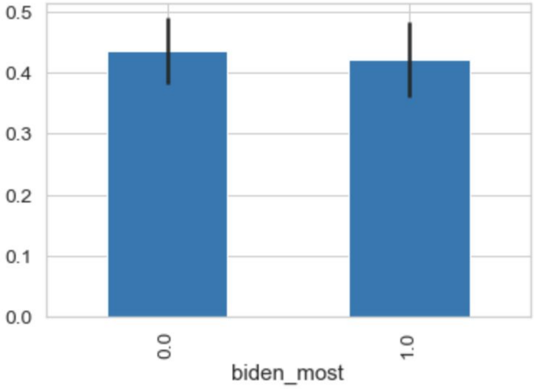
Since we were unable to obtain labeled data for training and testing, our team relied on a package called TextBlob, that uses its own knowledge of words and sentiment to calculate two scores for a string:

1. Polarity - ranging from [-1.0,1.0] with -1.0 being completely negative and +1.0 being completely positive
2. Objectivity - ranging from [0.0,1.0] with 0.0 being completely objective and 1.0 being completely subjective

Polarity and Objectivity scores were obtained for each article based on the original (unprocessed) text strings. In order to determine if the New York Times treats the candidates differently, appearances of each candidate's name in an article were counted and if Biden's name appeared more than Sanders' then the article received a "1" value in the column called "biden_most"; if Sanders' appeared more than Bidens', the column received a "0" value; if mentions were equal, the article was removed from this analysis. In total, there were 724 articles analyzed, of which only 209 mentioned Biden more than Sanders.



Mean Polarity Score



Mean Subjectivity Score

Based on these charts, one can see that on average articles mentioning Joe Biden more than Bernie Sanders tended to be more negative (lower polarity score), while mean objectivity scores didn't differ and tended to be on the more objective side (less than 0.5).

Individual Contribution Analysis

In MongoDB, we stored the information about fundraising and delegates-won results for each candidate in different states. Using these data, we looked at how confident we can be if we predict whether a candidate can win a state by using two features as independent variables, number of unique contributors in the state and total amount of donation raised in the state.

After randomly splitting the training and testing data, the logistic regression model gave an accuracy of 0.571 for Biden data set and 0.778 for Sanders data set. We further aggregated these two data sets by calculating the difference of Biden's and Sanders' corresponding values for these two features. In other words, we got two new columns which indicate, for each state in the data set, how much more amount of donations Biden got than Sanders and how many more unique donors Biden got than Sanders. By doing the logistic regression again of these two new features on Biden's delegate-won results for each state, we got an accuracy of 0.667.

Implications and Recommendations

The topic modeling done on headlines and article text shows that there are clear differences in how Senator Sanders and Vice President Biden are written about in the New York Times. Senator Sanders' name appeared more often and in topics with more varied words, such as "Russia" and "Trump", while Biden was talked about almost exclusively in conjunction with other democrats.

The analysis on campaign contributions predicting state primary outcomes provided data-analytics backup for intuition that the candidate campaign teams can infer state delegates-won results from the fundraising information of the candidate and his/her competitor. They can know if the candidate is more likely to win the state by just looking at how many people contribute in the state and how much they donated already. However, one of the limitations here for this analysis is small sample data, considering not all of the states had the primary. Except for that, further improvement can be made if we can collect TV ad spendings, number of rallies held, number of direct mail pieces sent out, etc, and put these additional data into regression to get higher accuracy.

In early February, Senator Sanders likely received a boost in media attention due to his close second-place finish in the Iowa caucuses and his win in the New Hampshire primary. These results may have influenced the amount of New York Times coverage he was receiving during the time period analyzed, accounting for the large disparity in articles mentioning the Senator more than Vice President Biden. Given the fact that our team now knows that Joe Biden won South Carolina on February 29th by a large margin, leading to multiple candidates dropping out of the race and endorsing him ahead of Super Tuesday, it would be interesting to the campaign for us to analyze the differences in media coverage before and after the shift in votes.

Conclusion

Our project provided new ways of analyzing and predicting candidates' performances by using media data like journals to do topic modeling and sentiment analysis so as to grasp a better sense of the external factors' influence on a candidate's primary results.

Specifically, by looking at Joe Biden and Bernie Sanders, we found that New York Times on average mentioning Joe Biden more than Bernie Sanders, and tended to be more negative towards Biden. Additionally, the logistic regression is proved to be useful to get a relative accurate prediction on whether the candidate can win a state based on fundraising information.

Appendix

A. Topic Modeling with Headlines

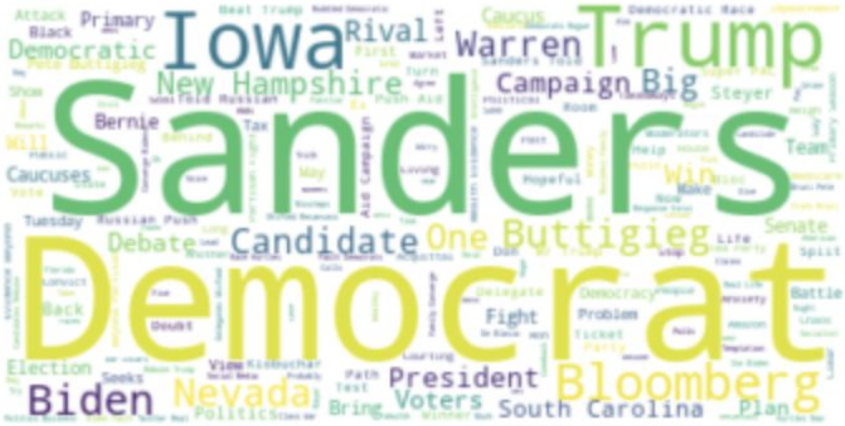
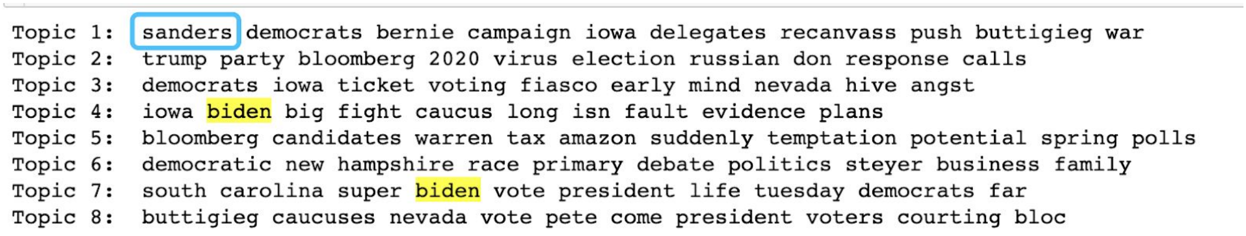
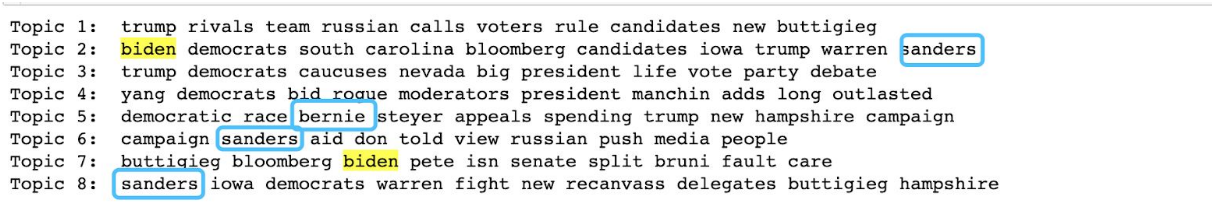


Figure 1. Word Cloud for NYT headlines.



```
Topic 1:  sanders democrats bernie campaign iowa delegates recanvass push buttigieg war
Topic 2:  trump party bloomberg 2020 virus election russian don response calls
Topic 3:  democrats iowa ticket voting fiasco early mind nevada hive angst
Topic 4:  iowa biden big fight caucus long isn fault evidence plans
Topic 5:  bloomberg candidates warren tax amazon suddenly temptation potential spring polls
Topic 6:  democratic new hampshire race primary debate politics steyer business family
Topic 7:  south carolina super biden vote president life tuesday democrats far
Topic 8:  buttigieg caucuses nevada vote pete come president voters courting bloc
```

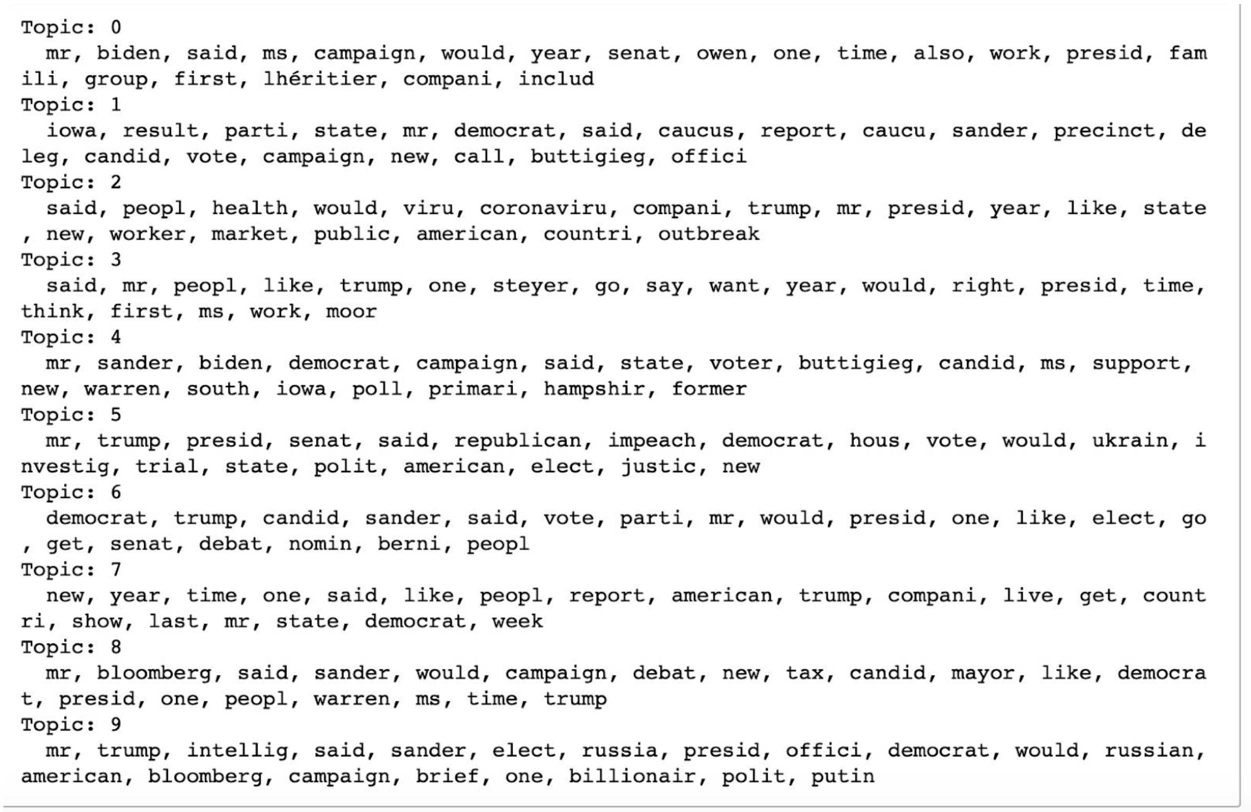
Figure 2. LSA Topics



```
Topic 1:  trump rivals team russian calls voters rule candidates new buttigieg
Topic 2:  biden democrats south carolina bloomberg candidates iowa trump warren sanders
Topic 3:  trump democrats caucuses nevada big president life vote party debate
Topic 4:  yang democrats bid rogue moderators president manchin adds long outlasted
Topic 5:  democratic race bernie steyer appeals spending trump new hampshire campaign
Topic 6:  campaign sanders aid don told view russian push media people
Topic 7:  buttigieg bloomberg biden pete isn senate split bruni fault care
Topic 8:  sanders iowa democrats warren fight new recanvass delegates buttigieg hampshire
```

Figure 3. LDA Topics

B. Topic Modeling for Article Body Text



```
Topic: 0
  mr, biden, said, ms, campaign, would, year, senat, owen, one, time, also, work, presid, fam
ili, group, first, lhéritier, compani, includ
Topic: 1
  iowa, result, parti, state, mr, democrat, said, caucus, report, caucu, sander, precinct, de
leg, candid, vote, campaign, new, call, buttigieg, offici
Topic: 2
  said, peopl, health, would, viru, coronaviru, compani, trump, mr, presid, year, like, state
, new, worker, market, public, american, countri, outbreak
Topic: 3
  said, mr, peopl, like, trump, one, steyer, go, say, want, year, would, right, presid, time,
think, first, ms, work, moor
Topic: 4
  mr, sander, biden, democrat, campaign, said, state, voter, buttigieg, candid, ms, support,
new, warren, south, iowa, poll, primari, hampshir, former
Topic: 5
  mr, trump, presid, senat, said, republican, impeach, democrat, hous, vote, would, ukrain, i
nvestig, trial, state, polit, american, elect, justic, new
Topic: 6
  democrat, trump, candid, sander, said, vote, parti, mr, would, presid, one, like, elect, go
, get, senat, debat, nomin, berni, peopl
Topic: 7
  new, year, time, one, said, like, peopl, report, american, trump, compani, live, get, count
ri, show, last, mr, state, democrat, week
Topic: 8
  mr, bloomberg, said, sander, would, campaign, debat, new, tax, candid, mayor, like, democra
t, presid, one, peopl, warren, ms, time, trump
Topic: 9
  mr, trump, intellig, said, sander, elect, russia, presid, offici, democrat, would, russian,
american, bloomberg, campaign, brief, one, billionair, polit, putin
```

Figure 1. Topics and their Top 20 words