Political News Analytical Toolkit

www.newsanalyticaltoolkit.com

Ian McCann
Galvanize

mcian91@gmail.com
github.com/iamianM/news_analysis

Motivations

The aim of this project was to **provide users with tools to examine political news and the outlets that write it.**

In a political climate growing ever more divided I think it is important for people to have the tools to compare news articles and organizations. I have developed tools that allow a user to **compare news sites by their mood, sentiment, and objectivity toward certain topics**.

A user can pick an analytical piece, opinionated piece, neither or both on a certain topic to read.

This allows one to be informed about the political topics in the news, how they are covered and in what way by different outlets.

The hope is that a user can view all sides of an issue and come to a conclusion on their own about how they feel towards it.

Data

The data used in this project was collected continuously over the past month from 12 news site's political RSS feeds. These feeds update each site's articles frequently.

The articles were preprocessed by removing stop-words, punctuation, quotes and tweets. Then, they were lemmatized, as well as, turned into bigrams, trigrams and quadgram. Finally, words that appeared in more than 50% of articles or less than 20 were removed.

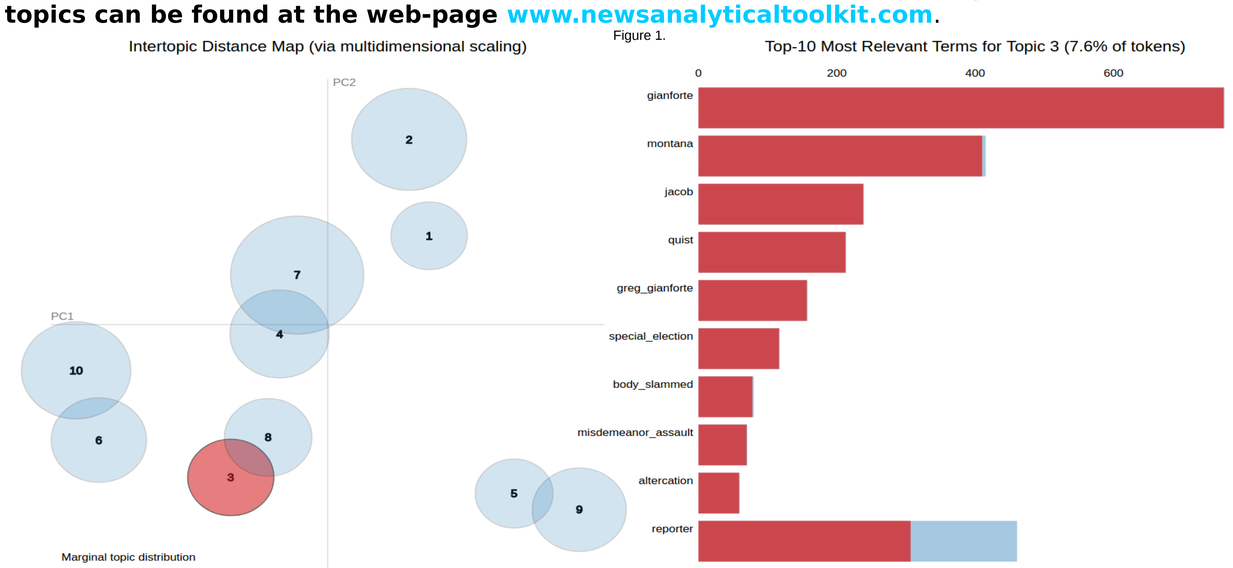
Topic Modeling

Latent Dirichlet Allocation (LDA) is a probabilistic method used to **discover latent topics** within series of documents and cluster them accordingly. LDA was used in this project to determine **55 topics discussed in over 7000 political articles from 12 news outlets**. Each news article can be considered a mix of multiple topics and LDA assigns a set of topics to each with a probability of it pertaining to that topic. Each topic has a set of words with probabilities of being related to it. Articles with a high frequency of words that have high probabilities of being in a topic will themselves have a high probability of being in that topic. The assumption is that the articles cover a small set of topics and the topics use a small set of words frequently.

Example of Topic Analysis

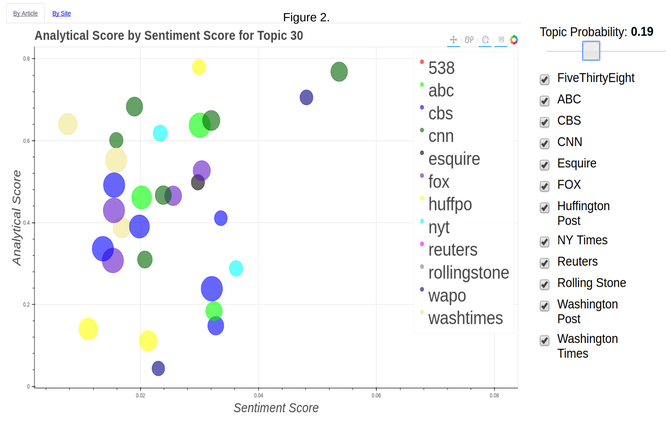
The left side of the figure below is an interactive visualization of the topics created using LDA. **Topic 30** was selected and on the right you can see the top 10 most relevant terms for that topic. This topic will be used as an example of the analysis I perform on all topics. The **analyses for all other topics can be found at the web-page www.newsanalyticaltoolkit.com**.

Figure 1.



Sentiment Analysis

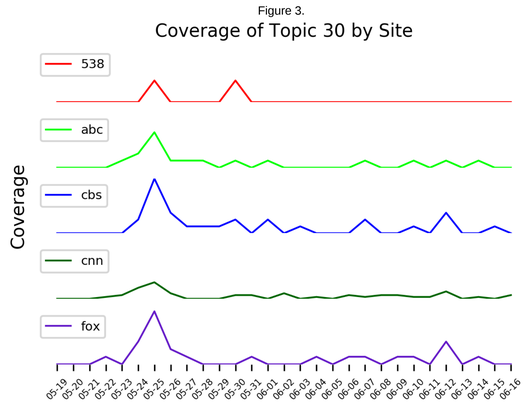
Figure 2.



The graph above plots all articles related to topic 30 by their **analytical score** versus their **sentiment score**. The analytical score was determined using IBM Watson's ToneAnalyzerV3 and the sentiment score was determined using the state-of-the-art sentiment analysis library SentiWordNet. The size of the circle indicates the probability the article relates to this topic and the slider on the top right allows you to set a probability threshold the article must reach to be shown. The check-boxes allow you to select which news sites you would like to view. You may also click any circle and it will open the article for you to read.

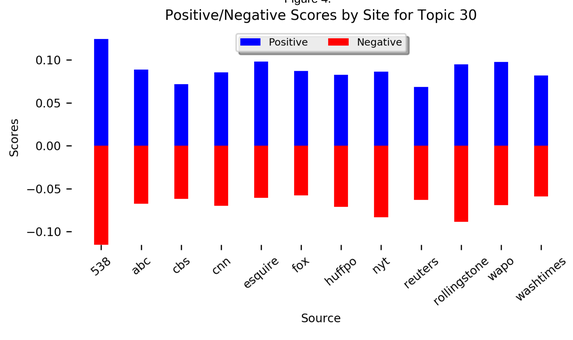
The following is a spark-chart of **coverage** for each date by news source.

Figure 3.



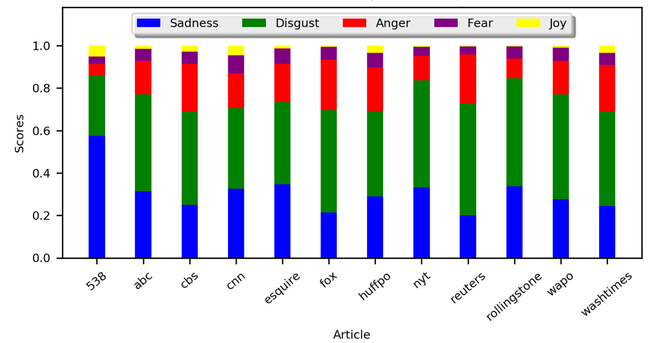
The following chart shows the average **positive and negative sentiment** of each site for topic 30.

Figure 4.



The following is a stacked bar chart of the five **emotional** scores ToneAnalyzerV3 calculated by each site.

Figure 5.



Conclusion

Topic 30 is about Greg Gianforte, who won the U.S. House special election on May 25th, 2017. The day before that he was accused of "body-slamming" a reporter and on June 12, 2017, he pleaded guilty to the assault charge.

If you look at the most relevant terms in figure 1, you can see many of the words I used to summarize this incident. In figure 3 you can see an up-tick in coverage around May 24th (the day of the assault) and June 12 (the day of the sentencing). In figure 5, we can see the moods that each site had toward this topic. The most common mood was "**Disgust**".

I believe this example demonstrates how LDA has successfully grouped these articles into the right topic and the sentiment analysis has correctly labeled how the news sources covered the topic.

A user can use these tools to examine which site covered this topic the most or the least, whether they discussed it in an analytical or subjective way, and how they relate to other news outlets considered in this project.

Libraries

1) Genism
2) PyLDAvis
3) IBM's Watson's ToneAnalyzerV3
4) SentiWordNet
5) Bokeh