Exploring Trending Topic Bias in News vs. Social Media

Jonathan Galsurkar (jfg2150), Moorissa Tjokro (mmt2167), Nitesh Surtani (ns3148)

1 Abstract / Introduction

Because of the constantly changing nature of the world and an increasing use of technology, having access to current events and up-to-date information has made people's lives more connected and visible in many ways. With social media growing as a platform to present information, it is crucial for news and media companies to bring accurate, unbiased, and edifying news both locally and around the world in a timely fashion. The news being talked about, however, may not represent what user of social media are talking about. Exploring biases between most trending topics on news and those being spoken about on social media would hence give both media companies and today's society the opportunity to observe the impact of information on people's interests and public opinions on news and various events throughout the world.

Currently, there is no specific approach that effectively captures how different topics that are trending on news platforms differ from those on social media. With events happening in distant locations that can impact economies, political tides, and various commodities over the long term, the development of an exploration pipeline is important to understand how and what kind of information is being listened to and learned from.

To maintain a quality standard of the accuracy and immediacy of news, the bias-trend platform can be further improved by incorporating a real-time feature and trend comparison analysis. The analysis would allow users to see the topic biases between two platforms, and the real-time feature would keep users informed on what is currently happening in the world and how it is perceived through the two platforms.

This problem is relevant to the course because of the sheer amounts of data that will be worked it, making distributed systems and cloud computing an integral part of the project's success.

2 One Sentence Summary

We will develop an exploration pipeline which allows a user to intuitively compare trending topics in news platforms versus those on social media platforms. We hypothesize that there is at least a moderate difference between topics presented in the news and topics spoken about on social media.

3 Audience and Needs

This project will prove to be a great asset to news, media, and any company that leverages consumer information. Companies will be able to see the differences between topics on the news versus topics actively spoken about on social media, allowing them to analyze the variance of discussions, understand what society cares about at the moment or in a given timeframe. News sources themselves can use our analysis and pipeline to figure out how to improve their presentation of certain topics to get more traction from readers as well as infer patterns from topics not gaining traction.

4 Data Preparation

4.1 Data Gathering

We used the New York Times API to extract the title, summary, and keywords of 7371 news articles from the last week of April. Next, we used the keywords gathered from the New York Times API to query and extract tweets using those keywords from the Twitter API. We gathered 33039 tweets from approximately the same date range.

4.2 Data Pre-processing

We concatenated the summary and title of the data provided from the New York Times API to form a sin-

gle news document. We treated each tweet as a single document as well. When reading the data we skipped any lines not properly encoded (utf-8). We also converted all text to lower case. Since most tweets only contain a few words, we removed any words that have less than 3 characters such as 'on', 'my', 'to', etc. These words would be popular but do not provide any value for making topics. We received much better results after doing this. Another common word we saw was 'https' and 'rt' which make sense since they refer to a link and retweet respectively. We got rid of word like this as well.

5 Approach

1. We started by extracting articles from the New York Times API and tweets from the Twitter API as discussed in Data Gathering.
2. We pre-processed the data as discussed in Data Pre-processing.
3. Our next step was the topic modeling. We used two methods to do so for both the News and Twitter data. These methods are LDA (Latent Dirchlet Allocation) and NMF (Non-negative Matrix Factorization).
4. With the prior knowledge that the news articles generally followed one of eight topics, we set our topic models to learn eight topics for each of the data sets.
5. Once each model learned the topics, we had a collection of distributions on words for each topic, allowing us to understand what each topic represented.
6. Having the distribution of words on all topics per data set, we now had the information to compare topics across the two data sets.
7. We analyzed the results and compared the topics themselves as well as the topic distributions for each data set.

6 (Best Case) Impact

We can show that there is a discrepancy between trending topics being shared and spoken about on social media versus those presented on news platforms. This will in turn allow for analysis on topics of human interest as well as analysis of topics of disinterest on news platforms.

7 Milestones

1. Gather news data into a structured format.
2. Gather social media data into a structured format.
3. Cluster similar data into topics based on platform using two topic modeling techniques.
4. Rank topics in clusters based on the probabilistic distributions obtained.
5. Use visualizations to see the distributions and to compare results across news/social media platforms.

8 Main Analysis

We use the Python library Scikit Learn to perform Topic Modeling using both Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) techniques. Using sklearn, we could include a seeding option which helped with the algorithm convergence in NMF and batch variants in LDA.

While LDA and NMF have differences in their mathematical underpinning, both algorithms return a collection of distribution on words (topic-word allocation) and the average distribution on topics across documents (document-topic allocation).

Another similarity is that both algorithms take as input a bag of words matrix. For instance, each document is represented as a row, with each column containing the count of words in the corpus. The output of both techniques are two small matrices: a document to topic matrix and a word to topic matrix, with their multiplication resulting in the bag of words matrix with the smallest squared error.

While both shares similarities, the main difference between these two topic modeling techniques is that LDA is based on probabilistic graphical modeling while NMF relies on linear algebra.

Using the feature extraction functionality in sklearn, we apply the tf-idf transformer to the bag of words matrix in NMF while we only use count vectorizer in LDA, which represents raw counts with both common and specified stop words.

8.1 Topic Modeling Using Latent Dirichlet Allocation

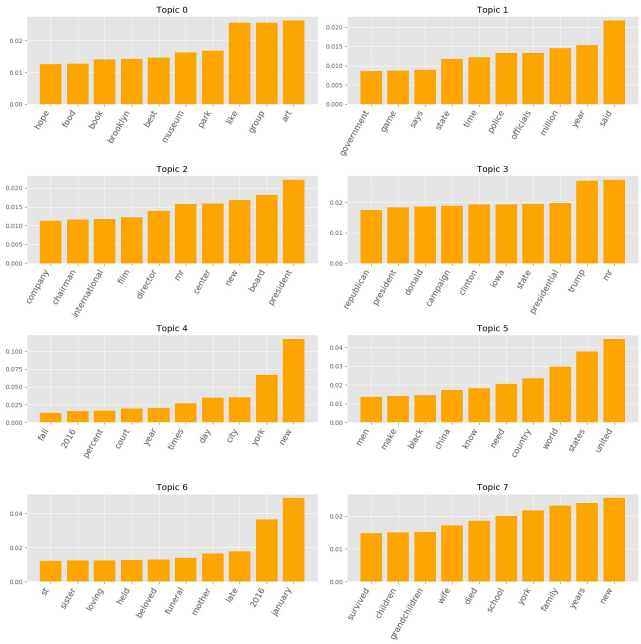

Figure 1: Top Words Associated with News Data Topics Learned using LDA

Using LDA, for the news data:

- Topic 0 seemed to represent the arts.
- Topic 1 seemed to represent States.
- Topic 2 seemed to represent Business
- Topic 3 seemed to represent Politics.
- Topic 4 seemed to represent New York.
- Topic 5 seemed to represent the World
- Topic 6 seemed to represent General .
- Topic 7 seemed to represent Life and Family.

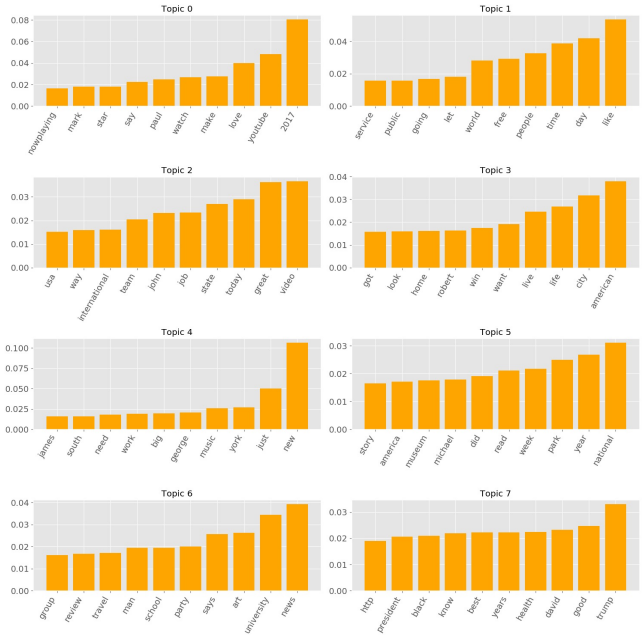

Figure 2: Top Words Associated with Twitter Data Topics Learned using LDA

Using LDA, for the twitter data:

- Topic 0 seemed to represent Entertainment.
- Topic 1 seemed to represent General.
- Topic 2 seemed to represent the World
- Topic 3 seemed to represent National.
- Topic 4 seemed to represent New York.
- Topic 5 seemed to represent the Arts.
- Topic 6 seemed to represent Entertainment .
- Topic 7 seemed to represent Politics.

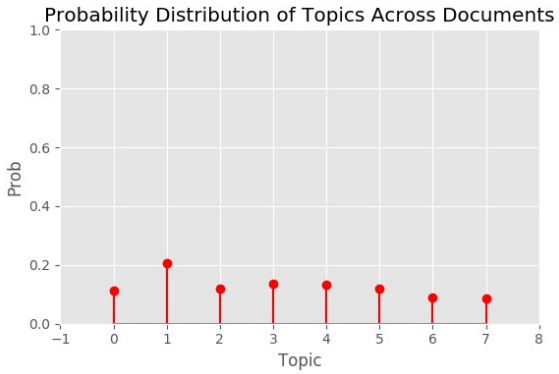

Figure 3: Distribution of Topics across all News Data using LDA

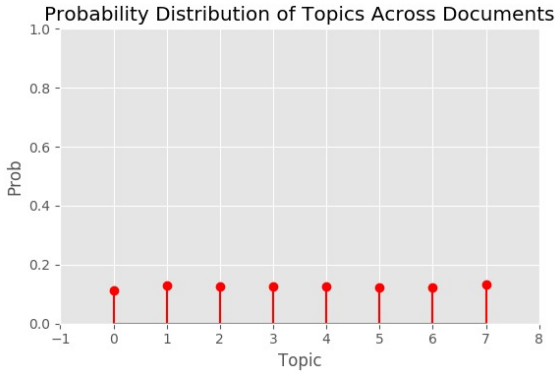

Figure 4: Distribution of Topics across all Twitter Data using LDA

Using LDA, we see that there is some overlap between the top topics spoken in New York times News and twitter. This overlap is mostly in Politics, the World, New York, and the Arts. Twitter seems to focus additionally on the topic of Entertainment and National information which we did not see in the New York Times while the New York Times had additional Topics not present in twitter such as Business and Life/Family. Some of this does intuitively make sense since we expect the news to focus on business while twitter may focus more on entertainment/pop culture.

8.2 Topic Modeling Using Non-Negative Matrix Factorization

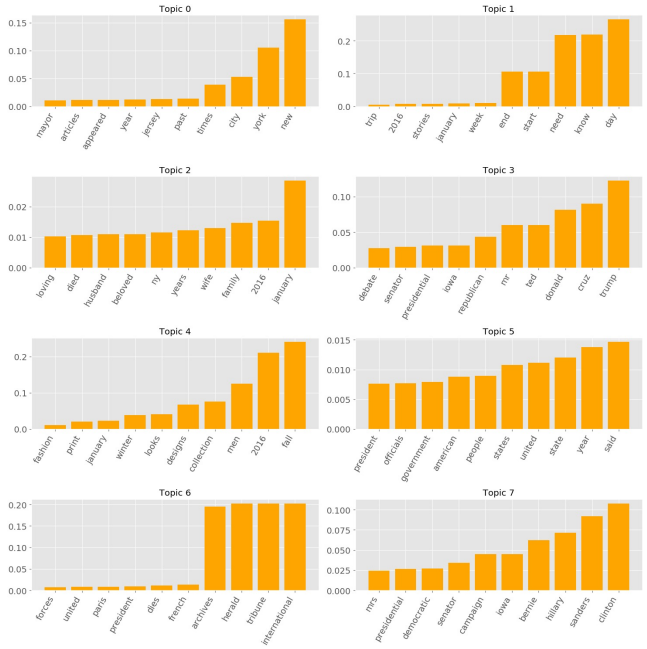

Figure 5: Top Words Associated with News Data Topics Learned using NMF

Using NMF, for the news data:

- Topic 0 seemed to represent New York.
- Topic 1 seemed to represent General.
- Topic 2 seemed to represent New Year
- Topic 3 seemed to represent Politics.
- Topic 4 seemed to represent Arts/Fashion.
- Topic 5 seemed to represent another Politics.
- Topic 6 seemed to represent the World .
- Topic 7 seemed to represent States.

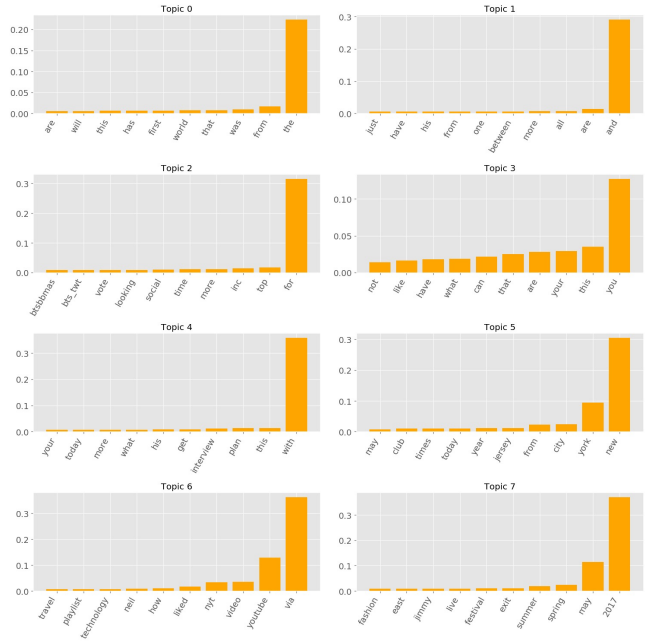

Figure 6: Top Words Associated with Twitter Data Topics Learned using NMF

NMF did not give interpretable topics for twitter. Therefore, it is hard to compare the actual results and differences between the platforms. We did notice that there were two politics related topics obtained from NMF, which can hint what the news was focusing on during the time the data set was acquired. We are confident that given more data, twitter would yield better results using this model. Given more time, we would not only gather more data, but work on training the model, which would in turn provide a more similar cluster of words per topic regardless of model.

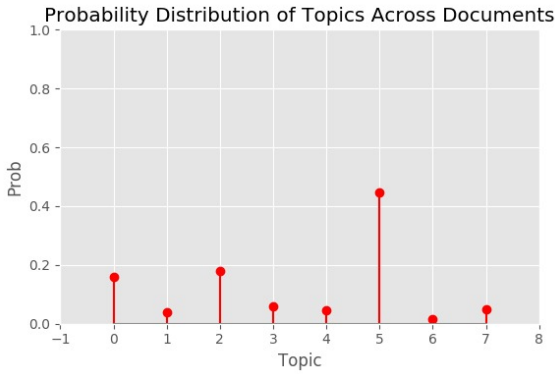

Figure 7: Distribution of Topics across all News Data using NMF

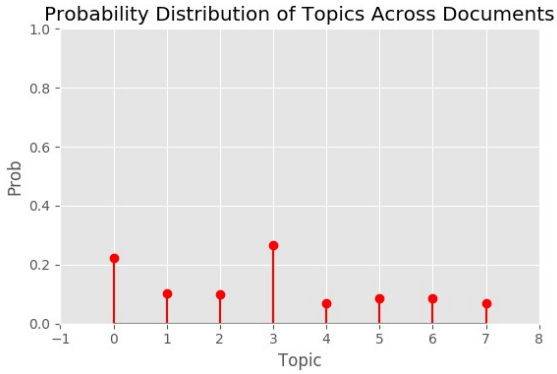

Figure 8: Distribution of Topics across all Twitter Data using NMF

It is difficult to compare the NMF results of the News and Twitter, however, we can see commonalities between the results of the NMF news and LDA news in terms of their trending topics such as Politics, New York, Arts/Fashion, and the World.

8.3 Results By Overall Topic Distributions Across All Documents

For each model, we ranked the topics across each data set using the mean of the probabilities of the topics across all document for that data set. For news data using LDA, the order of prevalence : States, Politics, New York, the World, Business, Arts, Death, Life and Family. For twitter data using LDA, there was seemingly no order of prevalence and they appeared to have an almost uniform probability.

For news data using NMF, the order of prevalence : Politics, New Year, New York, Politics2, States, Fashion, General, World. For twitter data using NMF, topics 3 and 0 had the highest probability, however, the topics didn't seem to provide any insights.

We see that the news focuses mostly on the States, Politics, and Business, while Twitter seems to evenly focus on all the topics. This intuitively makes sense since twitter is a social media platform specifically designed to speak a person's mind about anything they would want that is specific to themselves in that moment. The News on the other hand is geared towards providing specific information related to the latest information relevant for a majority of viewers.

9 Obstacles

9.1 Major obstacles

- We had a lot of difficulty gathering the data. The New York Times API would not provide full articles and only allow us to get recent articles by their title and summary. This also limited the amount of data we could use for the analysis.

- The Twitter API only allowed us to query tweets using keywords and at a rate of 180 per 15 minutes. The tweets also came from all across the world, being in various languages, etc. which made it more complicated to preprocess and model the data.

- In future work, assuming we have a way to store and access vast amount of the data that we gathered, we may not have access to the computing resources necessary to run real time analysis and visualizations on extremely large sets of data. We therefore did the analysis over smaller time periods, which could reduce the impact of the project, through showing a bias for even smaller time periods is still a contribution.

9.2 Minor obstacles

- In future work, we may not have access to the storage resources required to store years of news and social media data. One year of data alone can yield hundreds of gigabytes of data.

- With the API's used, we did not have access to data sets from certain social media or news platforms, thus possibly introducing slight bias to our results.

10 Conclusion

We have built a pipeline that will allow a user to analyze and compare news and social media data, however, they must gather the data themselves. With our analysis, we found that LDA provides better results than NMF.

We found that the news seems to be more reliable and provide more concrete topics than those in social media. With more data, a more concise and well rounded analysis can be done. The most challenging part of this project was the gathering of the data. Due to the limitations of the API's, our analysis was on a limited data set.

10.1 Moving Forward

We would create a script to constantly pull data from the API's over a longer period of time as well as gather many more keywords for extracting twitter data. We believe that with more data, we would have stronger results.

11 Additional Resources For Future Analysis

- Some computational time to run our optimizer algorithm to generate information from news and social media APIs.

- Access to a machine where we can install and run experiments, and possibly scale our system, using the current database prototype.

12 Literature Review

- *Background for the project:* This work does an extensive study on the Twitter trending topic [1] and studies the temporal behavior and user participation on these topics. They work on the complete Twitter data comprising of 41.7 million

user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. [4] works on exploring the popularity of news platform vs social media for consuming news.

- *Work the project relies and builds on:* [2] does some interesting work on finding the social media utterances that implicitly reference the given news articles. The first module of our system builds on the same approach on identifying the trending topics and querying Twitter to extract them.

- *Direct competitors:* We could not find any existing works which specifically focuses on exploring the relationship between news and social media to study the usage behavior of trending topics.

- *Alternatives to achieve the broader goal:* There is no relevant prior work done we found on this topic.

13 Acknowledgements

References

[1] Kwak, Haewoon and Lee, Changhyun and Park, Hosung and Moon, Sue, *What is Twitter, a social network or a news media?*, Proceedings of the 19th international conference on World wide web, 2010.

[2] Tsagkias, Manos and De Rijke, Maarten and Weerkamp, Wouter, *Linking online news and social media*, Proceedings of the fourth ACM international conference on Web search and data mining, 2011.

[3] Bandari, Roja and Asur, Sitaram and Huberman, Bernardo A, *The pulse of news in social media: Forecasting popularity*, arXiv preprint arXiv:1202.0332. 2012.

[4] Hermida, Alfred and Fletcher, Fred and Korell, Darryl and Logan, Donna, *Share, like, recommend: Decoding the social media news consumer*, Journalism Studies. 2012.

[5] Newman, Nic, William H. Dutton, and Grant Blank. *Social media in the changing ecology of news: The fourth and fifth estates in Britain.* International Journal of Internet Science 7.1 (2012): 6-22.

[6] Broersma, Marcel, and Todd Graham. *Social media as beat: Tweets as a news source during the 2010 British and Dutch elections.* Journalism Practice 6.3 (2012): 403-419.

[7] Stassen, Wilma. *Your news in 140 characters: exploring the role of social media in journalism.* Global Media Journal-African Edition 4.1 (2010): 116-131.

[8] Weeks, Brian E., and R. Lance Holbert. *Predicting dissemination of news content in social media a focus on reception, friending, and partisanship.* Journalism and Mass Communication Quarterly 90.2 (2013): 212-232.