Classifying Local News Television Transcripts Using RoBERTa

Team: Sam Wolken, Nick Pangakis, Chloe Ahn

1) Abstract

Local television news is one of Americans' most common news sources, but its content has received minimal study compared with its cable and digital counterparts. Relevant local information is necessary for citizens to cast informed votes in local elections. We amassed a novel dataset of approximately 18,000 closed captioning transcripts from local television news programs, which we divided into over 600,000 one-minute segments. We fine-tune seven RoBERTa models to detect common topics in local television news. These classifiers allow us to estimate the amount of attention devoted to different news topics by local television news programs and how the focus of local television news has changed over time. As far as we are aware, our project is the first to implement a transformer-based architecture (i.e., RoBERTa) to study local television news content. As such, our primary contributions are in application and data. Across seven classification topics, we find the series of RoBERTa models achieved an average precision score of 0.85, an average recall score of 0.876, and an average F1 score of 0.859.

2) Introduction

Local television news represents the only local news source that many Americans monitor. Despite being one of Americans' common news sources, its content has received minimal study compared with cable television and digital news sources. The stories covered in local news television are especially important in light of the recent trend towards nationalization in American political behavior and news coverage by U.S. outlets. As a result, the content that local television network affiliates choose to cover is the best opportunity for many Americans to get information about politics and policy outcomes at the state and local level, such as the local crime rate, the performance of public schools, how the local government responded to natural disasters, new legislation, the actions of political actors, and so on. Relevant local information is necessary for citizens to cast informed votes and hold politicians accountable.

What types of topics do local news stations cover? How much time do stations allocate to each topic and do patterns in coverage fluctuate over time? To study these questions, we amassed a novel dataset of approximately 18,000 closed captioning transcripts from local television news programs across three cities from 2014 to 2018 and created a series of classifiers to identify news topics in these transcripts over time. This project represents an initial step towards large-scale classification of local news transcripts and identifying which news topics these stations chose to cover and which they do not. We classify seven major topics covered by local news outlets: politics in general, national politics (i.e., the federal government), subnational politics (i.e., state and local government), crime, weather, sports, and disasters. We divided each television transcript into minute-long segments in order to classify local television news broadcasts at minute level. These minute segments are the input to our series of classifiers, and the outputs are predicted labels for each topic assigned to each minute-length segment. Our project's main outcome is a dataset that includes 667,573 minute-long segments classified along 7 binary categories. Because the data also include the date and the location of the recording, we can assess how local news coverage changes over time and across geographies.

To evaluate and train each classifier, we manually labeled a subset of news segments (indicating whether each news segment mentions any of the topics of interest). Then, we used cross-validation and tested the performance of our classifiers on held-out news segments. We measured performance using various metrics such as average precision, recall, and F1 score.

Across the seven topics, the RoBERTa models achieved an average precision score of 0.85, an average recall score of 0.876, and an average F1 score of 0.859.

3) Background

Initial work in natural language processing treated documents as unordered sets of words. As a result, early NLP methods (e.g., bag-of-words) used simplified representations of language that relied on relative term frequencies across and within documents. In recent years, however, researchers have begun training neural network classifiers to learn text representations and word embeddings. Prior work has also shown that pre-training models on a large corpus can significantly improve classification tasks. Some examples of this approach include Word2Vec (Mikolov et al. 2013) and ELMo (Peters et al. 2018). The most recent developments in natural language processing—especially BERT, RoBERTa, and other related techniques—feature pre-trained language models on significant amounts of unlabeled data. By using learned parameters from these pre-trained models, researchers can save significant time, learn complex word embeddings, and produce more accurate classification tasks. BERT was trained on BooksCorpus and Wikipedia data, which contains 800 million words and 2,500 million words, respectively. The key advantage of using BERT, RoBERTa, or another related technique is through implementing a strategy called transfer learning, which involves using the pre-trained model's parameters for task classification.

BERT and RoBERTa have been tested on newspaper articles and social media data. In academic literature, such analyses have primarily served as proof of concept or an opportunity to test different techniques against one another rather than for the analysis of media-related research questions. We build on prior work by introducing these techniques to a novel dataset drawn from a little-studied target population and by applying these techniques to gain leverage on substantive questions relevant to fields including political science and communication. We are drawing from the following articles to guide our approach to building the models:

1. How to Fine-Tune BERT for Text Classification?
    a. This paper includes specific guidance on fine-tuning BERT for classification and for text processing.
2. Fine-tune and host Hugging Face BERT models on Amazon SageMaker
    a. This article also offers guidance on implementing and fine-tuning BERT models in AWS SageMaker.
3. Moral Framing and Ideological Bias of News
    a. This project represents a conceptual model of sorts for our project. Using news stories from digital news outlets, the authors create fine-tune BERT classifiers for 12 different categories based on theories of media framing. We are focused on coverage of topics by local television stations, not how topics are framed, and we followed a similar approach using supervised classification with RoBERTa by creating a series of models.
4. Using Roberta classification head for fine-tuning a pre-trained model
    a. This blog post gives an excellent template for creating RoBERTa models. We heavily adapted our code from this template.

4) Summary of Our Contributions

1. **Contribution(s) in Code: N/A**
2. **Contribution(s) in Application:** Our project represents the first attempt we know of to classify local television news content using RoBERTa or a similar technique.
   This application of RoBERTa sheds light on the substance of an important news source for Americans with implications relevant to multiple academic fields.

3. **Contribution(s) in Data:** We generated a novel dataset of over 18,000 local television news program transcripts by scraping text from Archive.org's television repository. This particular dataset of local news television programs is unique to our project and, to our knowledge, no other projects have attempted to classify local television news coverage at this scale. Using the RoBERTa classifiers, we classified every minute of every transcript along 7 topic categories. The final outcome dataset is 667,573 minute-long local television news segments classified along the 7 categories.
4. **Contribution(s) in Algorithm: N/A or introduce in 2-3 sentences**
5. **Contribution(s) in Analysis: N/A or introduce in 2-3 sentences**

5) Detailed Description of Contributions

5.1 Methods

Because our primary contributions are novel data and application of ML techniques to a substantive question, we will focus on our data.

Data collection

Because we are using NLP techniques to analyze local television data, the primary initial task was to create a corpus of local television news transcripts. Archive.org maintains a library of over 2 million television transcripts that are generated from closed captioning. We found that researchers had previously written code for the general purpose of scraping this television transcript collection (https://github.com/notnews/archive_news_cc). We first created a list of network affiliate television stations in three major northeastern media markets: Boston, New York City, and Philadelphia. Then, we scaped all available transcripts from these stations. To narrow our data down to local television news programs, we reviewed the program titles available as metadata for each television program. We then used keyword searches to identify titles that were likely to be news programs for each channel. We confirmed these program titles through manual searches. Once we had a reliable keyword for at least one news program for every television station included in our sample, we subset the transcript dataset down to only local news programs. The resulting dataset has 17,904 transcripts from news broadcasts that aired between 2014 and 2018. The availability of data and our keyword approach to filtering the data resulted in a dataset that contains about 80% transcripts from the Philadelphia media market and 10% each from Boston and New York City. In the supplemental material, Figure 1a shows a distribution of the collected segments over time. The geographic imbalance in our sample reflects quirks of the data repository where it was scraped from, as Archive.org is a nonprofit organization and does not ensure balanced coverage of programs from each local television station.

Data cleaning and preprocessing

We combined the scraped transcripts into a CSV with a row for each unique news program. The columns included the transcript itself, a unique identifier for each transcript, channel name, show name, airing/running time, region/city of the broadcasting channel, and date. Using the baseline information provided from the original transcript files, we identified whether each transcript was from the closed caption transcript officially provided by each news channel or from the automatically generated one after the episode had aired. We manually compared whether the quality of automatic transcripts significantly differ from that of official transcripts. We did not detect a notable difference between the two, so we decided to include both. All of the transcripts included minute by minute timestamps embedded within the transcript (e.g., "5:57 amthey are all together. They are all part of what it's all about. … 5:59amtombstone is coming up new at …"). We divided each transcript into minute-long segments by splitting the transcript text with a string pattern (e.g., "[0-9]{1}:[0-9]{2} am"). We use minute-long segments as

the samples for our topic annotation and our models because they allow for more granularity in detecting how much attention news programs gave to certain topics. Each minute-long segment has approximately 130 words on average, not including stop words. Therefore, we believe the minute-long segments are appropriate input for our models. Past research has documented changes in preprocessing decisions can significantly alter the model performance and substantial findings from NLP (e.g., Camacho-Collados and Pilehvar, 2017). Currently, we have focused on a basic level of preprocessing, such as removing HTML/hex characters and unnecessary white space/lines and lowercasing.

Data annotation

To develop a labeled dataset for use in training, validation, and testing of our RoBERTa models, we had three researchers (the authors) read 100 randomly selected segments. Each of the researchers coded the same 100 segments on 9 different dimensions based on a pre-specified codebook. The codebook was designed to identify topics of theoretical importance and provide clear guidelines to increase intercoder reliability. These topics included politics, the economy, sports, crime, immigration, and the weather. Among these main topics, we also included a variety of subtopics including national politics, subnational politics, and humanitarian immigration. The codebook narrowly defined the type of information needed to operationalize each respective topic and gave several examples to assist coders. For every segment, coders were instructed to label each feature as 1 if the topic appeared in the segment and 0 if the topic did not appear. Table 1a in the supplemental material shows the intercoder reliability for each topic and the average frequency of each topic per segment. Intercoder reliability was measured with Krippendorff's alpha.

After the initial round of coding, we discussed confusing aspects of the original codebook and revised minor portions as needed. An alpha of 0.67 is regarded as an indication of a reliable category, so we considered categories that surpassed this threshold to be promising for supervised classification (Krippendorff 2009, 354). A main goal of our revisions to the initial codebook was to improve the reliability of our subnational politics category and to add an additional category: disaster. We also decided to drop economic news, immigration, and humanitarian immigration, due to their low frequency or low intercoder reliability. As a second round of coding, we coded another 150 randomly selected segments together and assessed intercoder reliability again. For this final round of coding, Table 1b in the supplemental material shows the intercoder reliability for each topic and the average frequency of each topic per segment. Importantly, each topic maintains an intercoder reliability above the conventional benchmark of .67. Following this final round of coding the same segments, we each separately coded about 300 segments along the 7 dimensions discussed above. In total, we had 1,017 minute-long segments coded along 7 dimensions. These 1,017 segments served as the inputs to our model.

Fine-tuning RoBERTa models

We used these 1,017 labeled posts to fine-tune RoBERTa seven models: one to classify each topic. We decided to specify seven distinct models, rather than a multiclass model, following the approach of others who have performed supervised classification of news content for a variety of labels that are not mutually exclusive [Mokhberian, 2020]. Our primary consideration was that transcripts were frequently labeled as containing multiple topics (e.g., national politics and crime, sports and weather, etc.), which is to be expected because each segment reflects one minute of airtime from the news program. Television news frequently touches on numerous topics quickly.

The one-minute transcript segments were tokenized using the pretrained base RoBERTa model that was developed by researchers at Facebook [Liu 2019]. We followed well-established guides for fine-tuning RoBERTa models in our decision-making process for preprocessing [Sun, 2019]. We chose a maximum sequence length of 512 characters. The models were fine-tuned

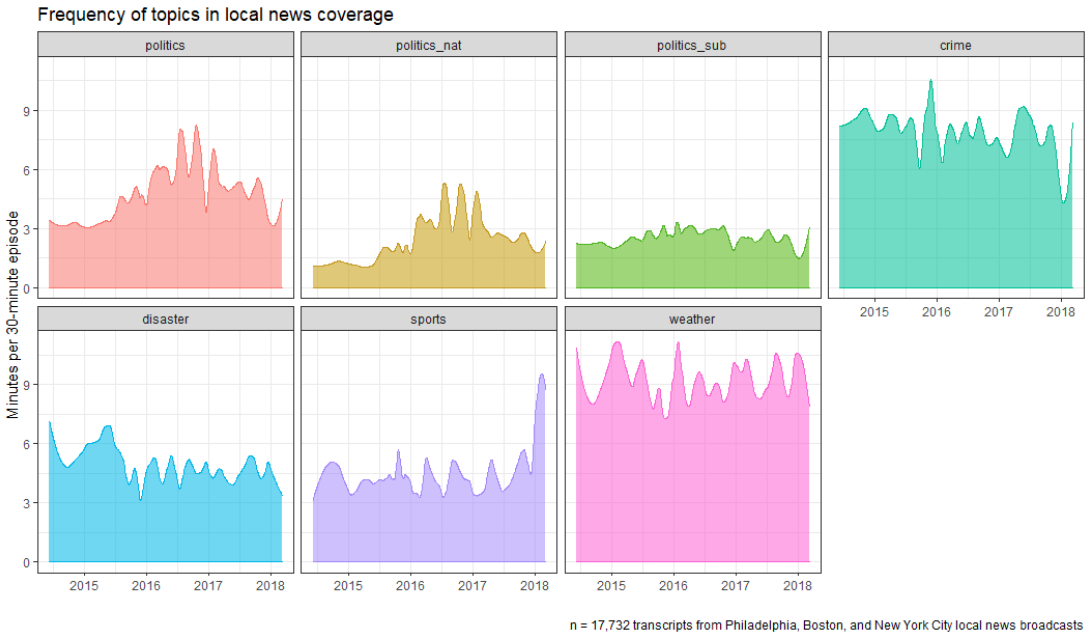
for 25 epochs with a learning rate of 0.00005. We used the AdamW algorithm as the optimizer; the primary advantage of AdamW over other optimizers is that it includes a moving average of the gradients and their squares and uses both to update the parameters. We used a batch size of 8 for both training and testing.

5.2 Experiments and Results

To measure the performance of our classifiers, we measured precision, recall, and F1 score. The table below shows these scores across the seven topics. These precision metrics were generated using approximately 200 held-out posts from the hand-annotated dataset of over 1,000 segments. Across the seven classification topics, our main finding is that the RoBERTa model achieved an average precision score of 0.85, an average recall score of 0.876, and an average F1 score of 0.859.

	Topic	P	R	F1
1	Politics	0.719	0.92	0.807
2	National Politics	0.812	0.929	0.867
3	Subnational Politics	0.8	0.857	0.828
4	Crime	0.935	0.811	0.869
5	Weather	0.942	0.925	0.933
6	Disaster	0.781	0.758	0.769
7	Sports	0.963	0.929	0.945

What types of topics do local news stations cover? How much time do stations allocate to each topic and do patterns in coverage fluctuate over time? The key outcomes of our research are estimates of how much airtime local television news programs devote to seven key topics from 2014 to 2018. To these questions, we summed the minute-length segments within each program. The programs tend to be 30 or 60 minutes, so the number of segments per episode, leading to a range of 30 to 60 segments per episode in most cases. To account for the variability in episode length, we normalized the program-level observations to 30 minutes (e.g., if 6 minute-length segments in a 60-minute program mentioned crime, then the same episode normalized to 30 minutes would have a value of 3 for crime). The below figure shows the distribution of time dedicated to each topic per episode over time. The y-axis can be interpreted as an approximate measure of minutes per episode spent on each topic.



These results lend considerable face validity to our classification process. Coverage of national politics increases around the 2016 election with three distinct spikes that are directly attributable to: a) Donald Trump becoming the Republican nominee in June 2016; b) the November general election; and c) the January Inauguration. Even more striking is the dramatic increase in sports coverage in 2018, which is likely driven by the Philadelphia Eagles winning the 2018 Super Bowl (since our 2018 data is only from Philadelphia). Figures 4a and 5a in the supplemental materials shows the trends in national politics and sports in greater detail.

Several other interesting observations are worth highlighting. First, crime and weather are by far the most popular topics. Figures 2a and 3a in the supplemental material examine whether coverage fluctuates seasonally. These supplemental figures provide suggestive evidence that crime coverage is fairly consistent and that weather coverage may increase in the winter months of January and February. Second, there is only slight evidence for nationalization in political content over time. Local political content stays fairly consistent over time rather than decreasing substantially. However, the limited timeframe and geographic scope of our data prevent us from testing this hypothesis directly, and we will pursue this topic in future research.

7) Compute/Other Resources Used

To implement RoBERTa, we used AWS Sagemaker to train our models and make predictions. We used the ml.p3.2xlarge instance type. We set our volume size to 6GB.

8) Conclusions

There are several important outcomes from this project. First, RoBERTa models are able to identify and classify topics in local news transcripts with a high degree of accuracy (i.e., an average precision score of 0.85, an average recall score of 0.876, and an average F1 score of 0.859). Manual review of predicted labels confirmed that the models worked with a high degree of accuracy and made very few mistakes. Second, the labels predicted by our RoBERTa models capture variation in news coverage resulting from current events, which helps to establish the validity of our predicted dataset. Together, these two takeaways demonstrate that supervised classification of local television news using a transformer-based architecture is a very promising avenue for subsequent research in the fields of political science and communication.

With the utility of this approach established, we intend to use the labeled data to investigate substantive questions about local television news coverage and politics. This project was motivated by a question about nationalization: Because Americans are so focused on national politics, does local television spend less time covering local politics? Although we only find subtle evidence of nationalization in political content over time, this subtle finding is plausibly because our data is geographically and temporally limited. We intend to scale up soon.

When this project is extended to more news stations across a larger time frame, the data will prove extremely valuable to other researchers. In addition to analyzing the tendency towards nationalization, other studies could assess whether there is a link between individual-level attitudes reported in surveys and information reported in the media. For example, does increased coverage of crime prompt citizens to express more concern about crime? In future work, we hope to incorporate data on voting and analyze the effects of local television on elections.

Ethical Considerations, and Broader Social and Environmental Impact:

This project has considerable potential social impact. It is important that voters have access to information about local politics and policy outcomes in order to hold local politicians accountable. By knowing what content local news stations are talking about, we can contribute to understanding how citizens make voting decisions. Because we are analyzing data from television stations and not individuals, we are not aware of any potential ethical concerns.

Supplemental Material:

Figure 1a:

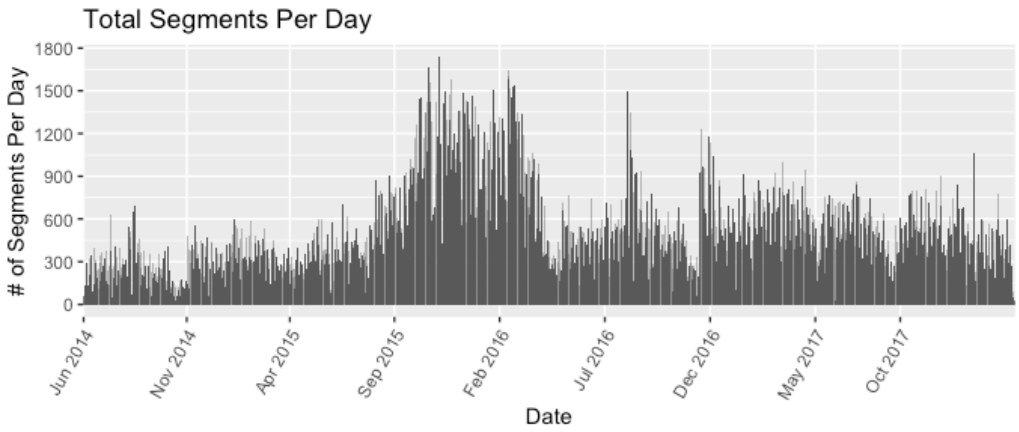


Table 1a:

	Topic	Avg. per Segment	Reliability (alpha)
1	politics	0.22	0.884
2	politics_national	0.15	0.869
3	politics_subnational	0.1	0.63
4	economic	0.017	0.391
5	crime	0.363	0.899
6	weather	0.267	0.813
7	sports	0.103	0.82
8	immigration	0.03	1
9	immigration_humanitarian	0.027	0.872

Table 2a:

	Topic	Avg. per Segment	Reliability (alpha)
1	politics	0.213	0.802
2	politics_national	0.133	0.923
3	politics_subnational	0.118	0.68
4	crime	0.287	0.826
5	disaster.emergency	0.14	0.705
6	weather	0.251	0.894
7	sports	0.129	0.842

Figure 2a:

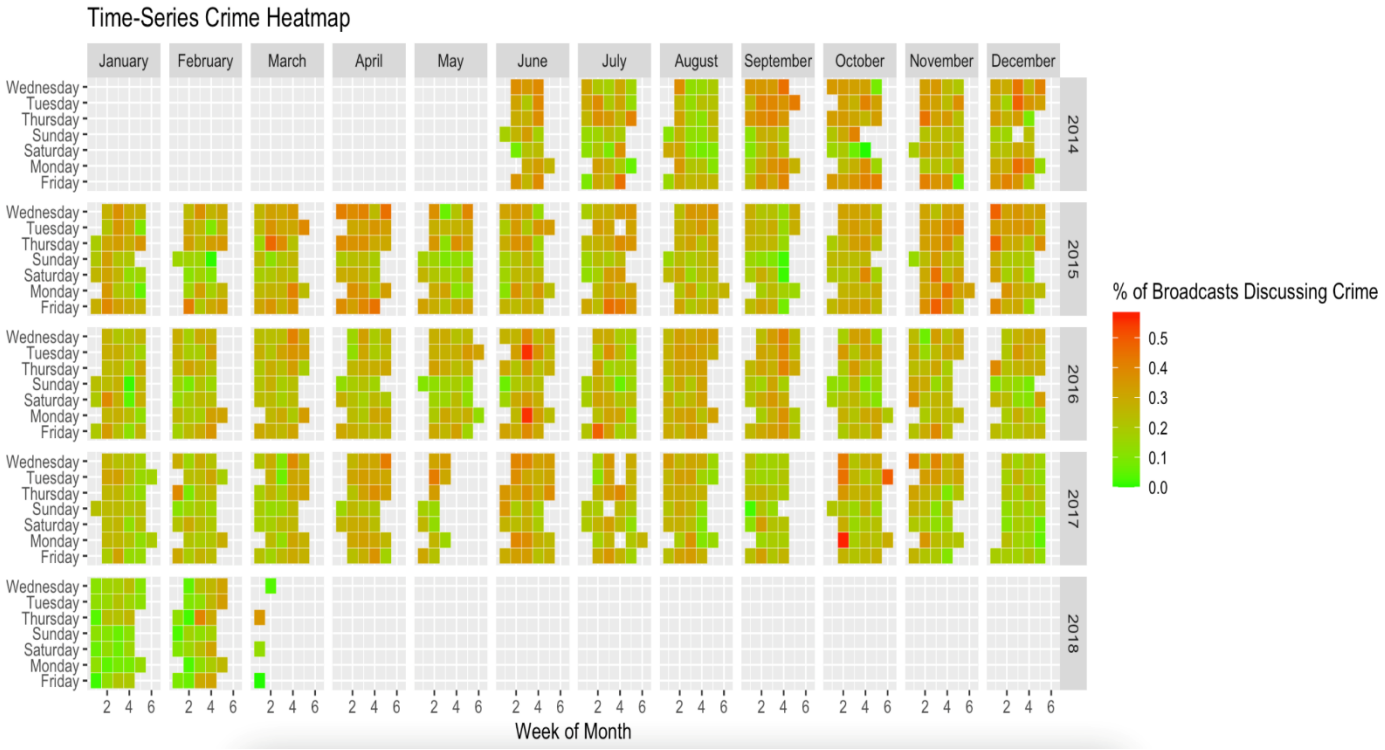

Time-Series Crime Heatmap
January
February
March
April
May
June
July
August
September
October
November
December
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
2014
2015
2016
2017
2018
2 4 6
Week of Month
% of Broadcasts Discussing Crime
0.5
0.4
0.3
0.2
0.1
0.0

Figure 3a

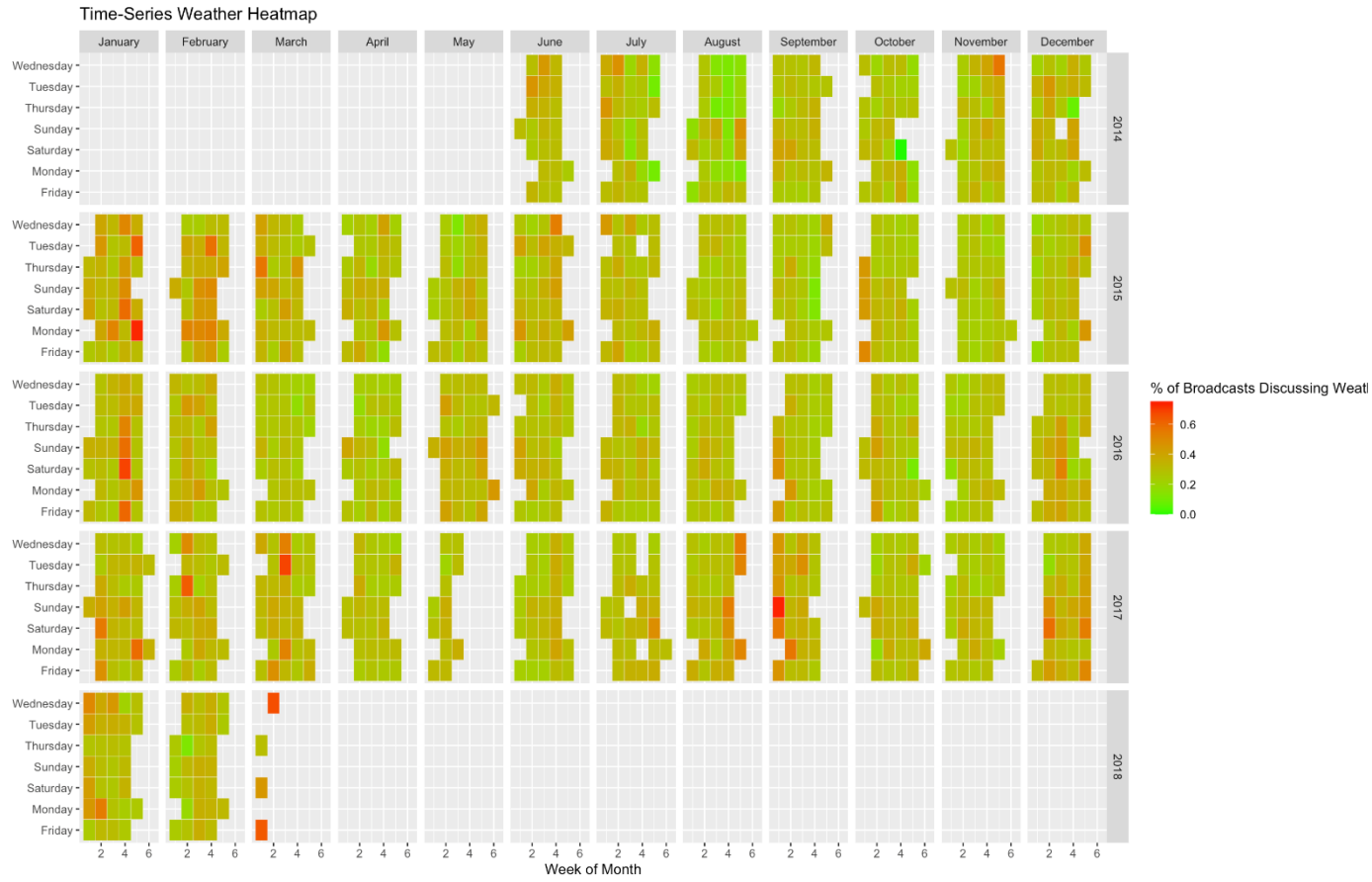
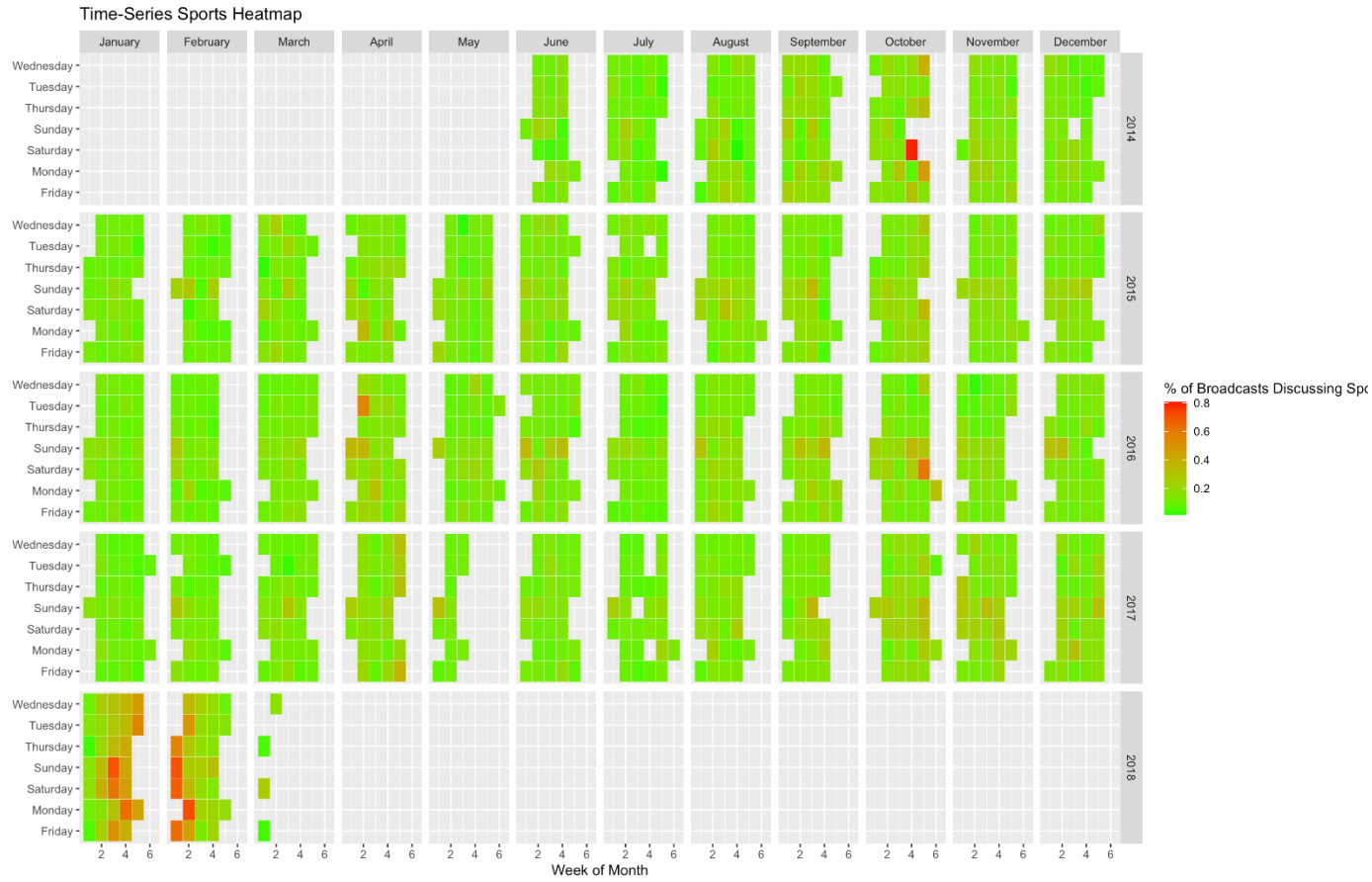

Time-Series Weather Heatmap
January
February
March
April
May
June
July
August
September
October
November
December
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
2014
2015
2016
2017
2018
2
4
6
Week of Month
% of Broadcasts Discussing Weat
0.6
0.4
0.2
0.0

Figure 4a:


Time-Series National Politics Heatmap
January
February
March
April
May
June
July
August
September
October
November
December
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
2014
2015
2016
2017
2018
2 4 6
Week of Month
% of Broadcasts Discussing National Pol
0.5
0.4
0.3
0.2
0.1
0.0

Figure 5a:


Time-Series Sports Heatmap
January
February
March
April
May
June
July
August
September
October
November
December
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
2014
2015
2016
2017
2018
2
4
6
Week of Month
% of Broadcasts Discussing Spo
0.8
0.6
0.4
0.2