TransciptCOVID:
A COVID-19 Transcript Analysis Across Countries

DS-GA 1015: Text as Data

Angela Marie Teng (at2507), Lakshmi S. Menon (lsm454)
Center for Data Science
New York University

Abstract

Despite the importance of using unsupervised techniques in predicting the outcome and efficacy of public health-related policies, the swift onset of the coronavirus pandemic has created little opportunity for researchers to study the topics of political texts. This paper demonstrates an application of structural topic models and latent dirichlet analysis on a corpus of coronavirus-related presidential transcripts. The main objective is to explore the relation between political sentiment and press briefings across the research sample of five countries, if any relationship exists. The researchers aim to answer the following question: "How do the ways through which country leaders engage with their citizens and address coronavirus-related topics differ? And how are these political ideologies reflected in the topics discussed by leaders in each transcript?"

Word Count: 2705 (excluding appendices, references, and footnotes)

Introduction

The coronavirus pandemic has affected the entire world, resulting in thousands of deaths, and a nearly global shut down. Countries have been praised and criticized for their varied responses to control the spread of the virus. While virus containment and public safety have been the primary goals for most, if not all, countries, the means by which they strive for these objectives are affected by their socio-political structure, national culture, and medical resources. We aim to study the political responses of five different countries by examining COVID-19 press transcripts. We hope to shed light on the effects that political climate could have on public health emergency responses.

Related Literature

Recent news reports have shed light on the varied approaches policymakers have taken to contain and handle the coronavirus pandemic. However, there have been discussions across news and social platforms that describe how some countries' policies towards dealing with COVID-19 are more effective than others. Particularly, the countries that have advocated earlier on for mass

testing, social distancing, and community quarantines have been lauded[1]. Specifically, there are three main ways through which citizens can get tested for coronavirus: polymerase chain reaction tests, antibody tests, and antigen tests[2]. Although these three tests are not equally effective, the countries which had higher rates of overall mass testing were more effective in containing the virus and were more successful at flattening the so-called "curve"[3]. For example, South Korea has seen success with this method, and as of May 1, 2020, the country experienced its first day without a new local case.

At the time of this writing, a number of studies have been influential to our approach and methodology. Kabir and Madria[4] worked on a similar analysis of COVID-19 related tweets, and used topic modeling to study "subjectivity and to model the human emotions during the COVID-19 pandemic". Additionally, Liu et al. did a similar study that applied digital topic modeling to news media outlets in the early stages of the virus[5]. Dong et al. also worked on COVID-19 topic modeling to better understand coronavirus research "hotspots"[6].

Theory and Hypotheses

Country leaders have been holding press briefings to provide updates to their nations about coronavirus spread and response. We expect that different leaders will have different priorities, based on the extent of the virus in that country, as well as their current national concerns[7]. Through this analysis, we postulate that these differences in ideals will be reflected in the way leaders address their nations, with respect to the topics they choose to discuss and the terms they use to discuss them.

---

[1] https://www.sciencemag.org/news/2020/03/mass-testing-school-closings-lockdowns-countries-pick-tactics-war-against-coronavirus

[2] https://www.cnn.com/2020/04/28/us/coronavirus-testing-pcr-antigen-antibody/index.html

[3] https://abcnews.go.com/Health/trust-testing-tracing-south-korea-succeeded-us-stumbled/story?id=70433504

[4] https://arxiv.org/pdf/2004.13932.pdf

[5] https://www.jmir.org/2020/4/e19118/pdf

[6] https://www.medrxiv.org/content/10.1101/2020.03.26.20044164v2

[7] Although these briefings may have the same intention of tracking the developing pandemic, we expect that leaders will discuss these goals differently.

Data and Models

To test this hypothesis, transcript data was collected from a sample of five countries, namely; the United States, the United Kingdom, New Zealand, Canada, and Australia. Topic models and emotion metrics were used to compare speech content between countries. To quantify the efficacy of each country in dealing with the coronavirus pandemic, we referred to two global coronavirus trackers; the Johns Hopkins Coronavirus Resource Center[8], and the New York Times Coronavirus Map[9]. We then compared the per capita metrics for each country to determine how effective each country's policies were in mitigating the spread of the pandemic.

Data Collection

Data was scraped from the records and transcript text on *Rev.com,*[7] a transcription service. We identified transcripts related to coronavirus based on the website's classification, as *Rev.com* has a dedicated page for '*Coronavirus Briefing & Press Conference Transcripts*'[10]. We limited our analysis to five countries: United States, Canada, United Kingdom, Australia, and New Zealand[11]. In order to maintain consistency among countries, we used only leader speeches, and thus did not include conferences by US State governors and mayors, as well as briefings by the World Health Organization[12]. After taking that subset of the texts, we were left with 120 articles. The distribution by country is shown below:

Country	Australia	Canada	New Zealand	United Kingdom	United States	**Total**

[8] https://coronavirus.jhu.edu/map.html

[9] https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html

[10] https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts

[11] This decision was made based on the texts available on the website. We first scraped the URLs and titles of all articles available under the coronavirus category, then narrowed them down to country leaders' speeches by checking for either a country name or leader name in the title. The processed data is saved as both a dataframe and as a document frequency matrix. We additionally converted the text columns from a list to a character vector, and changed its encoding from UTF-8 to ASCII, as the data contained UTF-8 quotations that needed to be removed in the preprocessing stage.

[12] We did, however, include briefings by the 'Coronavirus Task Force' in the case of the United States, as these often included the US President Donald Trump, and hence reflected the country leader's response.

Number of Transcripts	10	27	5	22	56	**120**

Table 1: Frequency of Transcripts by Country

In order to preserve context, we scraped each transcript as a single large text without segmenting it by speaker, as this would be in line with the goal of analyzing country response, as questions asked by reporters and responses by other speakers, such as health officials, still reflect the concerns and overall response of that country.

Data Description

We used articles from the first available transcript, on Feb 25, 2020, until April 23, 2020, the most current transcript on the date of web-scraping. We created a column for the country of each article, and populated it by matching keywords in the title. We also added another column with the date as a numeric object. The final dataset was a table with 120 rows of articles, and 6 columns, containing the URL, title, string date, numeric date, country, and text for each article.

url	title	text	date	country
https://www.rev.com/blog/tr...	Donald Trump Coronavirus Press Conference Transcri...	Thank you very much. Later this evening, we expect t...	Apr 23, 2020	US
https://www.rev.com/blog/tr...	Justin Trudeau Canada COVID-19 Press Conference A...	Today at sundown marks the beginning of Ramadan. ...	Apr 23, 2020	CA
https://www.rev.com/blog/tr...	United Kingdom Coronavirus Briefing Transcript April ...	138,078 people have tested positive for the virus. Th...	Apr 23, 2020	UK
https://www.rev.com/blog/tr...	Donald Trump Coronavirus Press Conference Transcri...	(silence)., Thank you very much. Appreciate it., A lot ...	Apr 22, 2020	US
https://www.rev.com/blog/tr...	New Zealand COVID-19 Briefing Transcript April 22	[foreign language 00:00:15] Welcome to day 28 of lev...	Apr 22, 2020	NZ
https://www.rev.com/blog/tr...	United Kingdom Coronavirus Briefing Transcript April ...	... and that whole MOD led by defense secretary Ben ...	Apr 22, 2020	UK
https://www.rev.com/blog/tr...	Justin Trudeau Canada COVID-19 Press Conference A...	[foreign language 00:00:08]. Before I get started, I wa...	Apr 22, 2020	CA
https://www.rev.com/blog/tr...	Donald Trump Coronavirus Press Conference Transcri...	Thank you very much everyone. Very big day today. I ...	Apr 21, 2020	US
https://www.rev.com/blog/tr...	New Zealand COVID-19 Briefing Transcript April 21	All right. Good afternoon everybody. [inaudible 00:00...	Apr 21, 2020	NZ
https://www.rev.com/blog/tr...	United Kingdom Coronavirus Briefing Transcript April ...	Good afternoon, and welcome to the Downing Street ...	Apr 21, 2020	UK
https://www.rev.com/blog/tr...	PM Scott Morrison COVID-19 Briefing Australia April 21	Welcome low levels. And we want to thank all Australi...	Apr 21, 2020	AU
https://www.rev.com/blog/tr...	Justin Trudeau Canada COVID-19 Press Conference A...	[Foreign language 00:00:08]. Hello, everyone. I want t...	Apr 21, 2020	CA
https://www.rev.com/blog/tr...	Donald Trump Coronavirus Press Conference Transcri...	Thank you very much everyone. Thank you., Followin...	Apr 20, 2020	US
https://www.rev.com/blog/tr...	United Kingdom Coronavirus Briefing Transcript April ...	Good evening from Downing Street where I'm joined ...	Apr 20, 2020	UK

Table 2: A Sample of The Transcript Data

Data Processing

To preprocess the data, we used natural language processing R libraries, including quanteda[13], topicmodels[14], stm[15], and readtext[16]. We utilized the approach discussed by Roberts,

[13] https://quanteda.io/

[14] https://cran.r-project.org/web/packages/topicmodels/index.html

[15] https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

[16] https://cran.r-project.org/web/packages/readtext/readtext.pdf

Stewart, and Tingley[17], and applied the the following preprocessing settings: conversion to lowercase, word stemming, removal of English stopwords, in addition to key words specific to the dataset[18], removal of punctuation, numbers, URLs, and special characters. We decided to remove numbers since our main goal was to identify topics discussed in each transcript. Given that numbers added no additional contextual insight, and there exists a separate document variable (docvar) that contained information about the date, we decided to drop all numeric characters from the corpus, before processing it into a document feature matrix.

We ended up with a document-feature matrix (DFM) comprising 120 documents, 8,791 features, and with 89.6% sparsity. Additionally, to reduce sparsity, we removed words that occured fewer than 20 times, and in fewer than 20 documents. The final trimmed document feature matrix had a sparsity level of 58.7%.

Modeling

A number of methods were used to model the topics discussed in each country's coronavirus transcript.

A. **Latent Dirichlet Allocation Topic Modeling**

Latent Dirichlet Allocation (LDA) is a "generative probabilistic model for collections of discrete data" including text corpora[19]. Using this method[20], we were able to determine the most likely topics that were being discussed for each country. In particular, we applied an LDA Gibbs model, with 3000 iterations and with k = 25 topics. The top 20 terms for the first 10 topics can be seen in the table below:

[17] https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

[18] We removed words like "speaker", "inaudible", and "foreign language", as these provided no additional information, but appeared frequently in the text corpus because of the nature of the transcription documents.

[19] http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[20] As described by Blei, Ng, and Jordan, LDA uses empirical Bayesian models to describe each item in a collection as a "finite mixture over an underlying set of topics", where each topic is "modeled as an infinite mixture over an underlying set of topic probabilities".

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
new	go	go	presid	will	will	governor	countri	need	know
state	test	presid	american	need	busi	china	go	make	countri
hospit	state	work	test	time	support	done	american	will	ship
see	phase	want	coronavirus	number	job	countri	new	money	will
york	back	great	state	take	govern	million	see	call	make
just	governor	thank	work	thing	next	deal	day	ask	nation
ventil	want	peopl	go	can	time	ventil	us	busi	year
governor	open	time	will	measur	today	open	health	lot	border
know	abl	mr	just	term	econom	job	look	sure	deal
need	care	lot	want	make	take	state	time	state	use
million	capac	like	dr	got	help	said	case	want	deploy
want	reopen	american	vice	week	protect	billion	great	data	mexico
said	will	pleas	let	death	economi	want	coronavirus	talk	drug
test	guidelin	come	hous	minist	announc	good	state	order	forc
week	new	virus	import	work	also	make	world	can	oper
case	come	crosstalk	said	point	employ	use	now	program	call
dr	million	take	everi	first	peopl	need	thank	small	question
will	talk	say	today	patrick	provid	know	today	well	back
mean	slide	job	public	keep	can	great	talk	ventil	fight
area	communiti	will	health	come	part	year	continu	week	militari

Table 3: Top 20 LDA Topic Terms (Topics 1 through 10)

The relative contribution of each LDA topic to each country's total topics was also calculated.

Country	13: The People	15: Work	25: Social Distancing	2: Reopening	24: COVID-19 Cases
AU	0.06422764	0.007429363	0.01560783	0.011050606	0.037381814
CA	0.05370287	0.416149126	0.01825687	0.011257352	0.017469405
country	0.10396843	0.003421615	0.03103983	0.005450437	0.018687340
NZ	0.09054348	0.007365489	0.08303024	0.010427168	0.407905730
UK	0.11789446	0.008039808	0.05805171	0.011950602	0.016733816
US	0.17446909	0.005563630	0.02868214	0.033293505	0.006057902

Table 4: Relative Contribution of each Topic by Country[21]

Interestingly, Table 4 shows that New Zealand's transcripts focused on COVID-19 cases (level of cases, number of cases), Australia's focused on the people (needs of people, empathizing with the audience), and Canada's focused on work (remote work, continuation of work). The United Kingdom and United States also focused on the people. We also experimented with topic stability, by using different random seeds and testing other text preprocessing methods. Although these approaches yielded interesting findings, the LDA topic models were relatively stable[22].

B. **Structural Topic Model**

[21] Note that the category "country" refers to the US, although these were state-level addresses rather than country-level addresses, so for LDA, the team decided to keep these transcripts as a separate category.

[22] Additional information about topic stability as well as the other methods we tried can be found on our Github repo here: https://github.com/angelaaaateng/TaD_Final_Project

The structural topic model (STM) is a generative model that applies word counts to identify the topical prevalence and topical content of a given document[23]. We began with a baseline STM that resulted in a topic model with 103 topics, and an 8791 word dictionary.[24] For this first iteration, the we removed rare terms, an approach common in topic modelling[25], but eventually decided against it given the relatively small size of the corpus, particularly when considering the number of documents per country.

To build an STM comparable to the LDA topic model, the data was processed using the recommended *textProcessor* function of the stm package[26]. Similar to the preprocessing applied on the LDA topic model DFM, the *textProcessor* function conducts basic preprocessing, including converting to lowercase, removing numbers, punctuation, and stopwords, and stemming words.

For the final iteration of the STM, we additionally removed custom stopwords[27]. These words included terms referring to the country or the people of that country, such as 'Australia' and 'Canadian', as well as pronouns that were common but did not convey much information regarding the topic, such as 'President'. The most common[28] words were also removed to reduce noise and increase accuracy.

The *searchK()* function was then used to select the optimal number of topics. Values between five and 20 were used first[29], and the optimal range was narrowed down to 10-15. The function was then run again in this range, and the optimal number of topics was chosen to be 10. With 10 topics, the model showed both high held-out likelihood and low residuals, with a threshold value for semantic coherence as well.

---

[23] https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

[24] For more details on this, see https://github.com/angelaaaateng/TaD_Final_Project/blob/master/master.Rmd

[25] https://cbail.github.io/textasdata/topic-modeling/rmarkdown/Topic_Modeling.html

[26] Note that this method is equivalent to the preprocessing that was applied to the LDA topic model, just using a different package (stm rather than quanteda).

[27] This decision was made based on the baseline results seen without additional processing.

[28] Words that occur in at least 110 articles

[29] The range of options was limited to a small value due to the small size of the dataset, as a larger number of topics would have led to overfitting as well as overlapping topics.

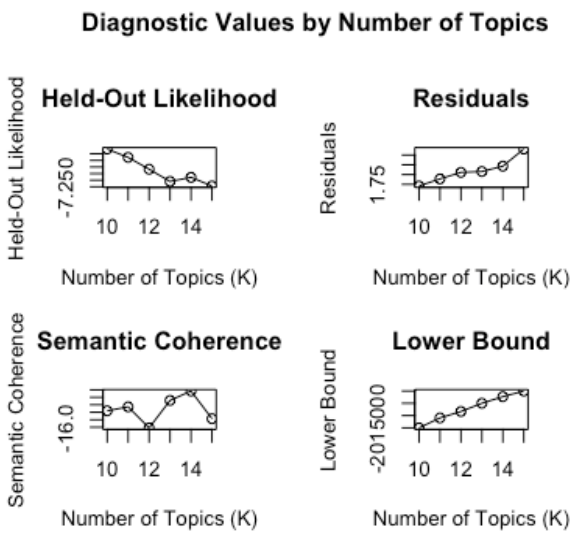


Table 5: Diagnostic Results produced by searchK() with K between 10 and 15

To see the effects of both country and date, the formula *~country + s(date)* was used for prevalence, and *~country* was used as the content parameter. The model converged after 25 iterations. A regression was also estimated on the trained model, using all ten topics and the same covariate equation[30].

C. **Latent Semantic Analysis**

To better understand lexical similarities between words and topics in the corpus, a latent semantic analysis (LSA) model was trained as well. The "tendency" of specific words to occur closely together and in similar contexts may reveal information about inherent and underlying relationships between those words. In this case, the corpus was subset according to country, and terms that were most closely related to "coronavirus" were examined based on each sub-corpus.

today	american	presid	work	home	can	us	just	peopl	mr
0.9765396	0.9752397	0.9726603	0.9724441	0.9717570	0.9691812	0.9691427	0.9680483	0.9679231	0.9669295

Table 6: Top 10 Closest Words to "coronavirus" in US Subset

measur	announc	sustain	area	individu	live	requir	like	ensur	make
0.9718580	0.9687324	0.9677074	0.9669484	0.9668320	0.9628756	0.9619457	0.9602210	0.9587614	0.9577827

Table 7: Top 10 Closest Words to "coronavirus" in AU Subset

[30] The results of these models are discussed in the results section along with some descriptive visualizations.

For example, for the 10 words that are associated with "coronavirus" in the US subset, we see that they are more vague, generally referencing the public, and are related to working from home. In Australia, on the other hand, there seems to be considerable effort made to quantify and measure the onset of the virus.

D. **Visualizing Emotion**

EmoLex[31] was used to visualize emotion between countries, following the methods used by Mesfin Gebeyaw in his article, *Parsing Text for Emotion Terms*[32]. The lexicon was then used to obtain the sentiment of each word, and the set of all tokens was filtered by removing words that had either positive or negative sentiment[33]. The lexicon includes eight different emotion categories: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Each of these categories was plotted over time, by their proportion in the whole corpus for each country[34].

Results

The covariate words produced by the *stm::labelTopics()* function are shown below. Understandably, most countries have many words referring to provinces or states within that country, possibly because the leader delivers updates on the situation in various areas during the briefings. For example, Australia has 'tasmania' and 'queensland', while the United States shows state names. It is also interesting to note the presence of the words 'Cuomo' and 'Mnuchin' in the US group. 'Cuomo', likely refers to New York State Governor Andrew Cuomo, who is known to often disagree with US President Donald Trump, most recently on issues of coronavirus response, while 'Mnuchin' refers to Secretary of Treasury Steve Mnuchin, who was involved in the logistics of delivering stimulus checks to support Americans. [13, 14]

[31] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[32] https://datascienceplus.com/parsing-text-for-emotion-terms-analysis-visualization-using-r/

[33] For this analysis, the full corpus of 120 articles was split into six separate dataframes by country, and the same analysis process was conducted with each set separately. The original text was used for each set, with only numbers and English stopwords removed.

[34] These plots, as well as a comparison between countries, are shown in the results sections.

```
Covariate Words:
Group AU: tasmania, cabinet, ahppc, queensland, victoria, hppc, hairdress
Group CA: olivia, french, stefanovich, cbc, tam, ctv, canada
Group NZ: bloomfield, zealand, dhb, wellington, tairawh, middlemor, northland
Group UK: nhs, england, beth, sky, pound, exit, london
Group US: congress, cuomo, popular, texa, appreci, florida, mnuchin
```

Table 8: Covariate Words Produced by STM

Additionally, the indicative words for each topic are shown in the appendix. The extensive presence of words such as 'hospit[al]', 'worker', 'ventil[ate/ator]', 'job', and 'safe' show that the country leaders overall speak with a focus on the well-being of their people, as well as the economy.

Topic 1: state, worker, school, lot, question, great, minist, french, busi, measur, realli, job, happen, sure, use, hospit, crosstalk, mask, quick, suppli
Topic 2: vaccin, realli, lot, covid-, question, french, year, develop, state, use, sure, case, million, great, drug, provinc, happen, got, food, call
Topic 3: cabinet, professor, measur, state, coronavirus, question, minist, foreign, travel, realli, busi, ensur, sure, let, communiti, languag, famili, case, prime, safe
Topic 4: state, question, realli, hospit, lot, case, governor, sure, great, job, call, happen, covid-, worker, use, okay, million, ventil, suppli, across
Topic 5: state, question, realli, lot, million, busi, case, worker, sure, job, happen, great, across, minist, coronavirus, call, communiti, there, move, measur
Topic 6: state, question, lot, realli, busi, hospit, worker, use, job, great, case, ventil, governor, happen, sure, million, suppli, across, minist, call
Topic 7: realli, question, lot, state, busi, pleas, close, measur, job, social, happen, sure, compani, great, worker, hospit, minist, distanc, someth, open
Topic 8: state, realli, case, lot, governor, question, great, happen, million, capac, sure, use, measur, communiti, across, job, there, phase, got, point
Topic 9: yvonn, state, case, nhs, travel, french, border, realli, question, coronavirus, measur, respons, spread, unit, china, global, communiti, busi, sure, use
Topic 10: state, question, data, lot, sure, suppli, happen, busi, realli, case, order, hospit, ensur, use, inform, crosstalk, call, measur, clear, there

Table 9: Indicative Terms for Each Topic

However, even when the countries speak on the same topic, the words they use and the specific aspects they focus on differ from each other. The comparative plots of words used by different countries within a given topic show that the UK and New Zealand both seem to focus more on the virus and its impact, with words such as 'vaccin[e]', 'case', 'data', 'death', and 'infect'. However, for both of these topics, the US uses more vague terms, such as 'great, 'sure,

and ‘lot’, as well as political terms such as ‘state’ and ‘governor’. In a similar vein, this sentiment is reflected in the words associated with “coronavirus” when the *associate()* function is applied on the LDA model.

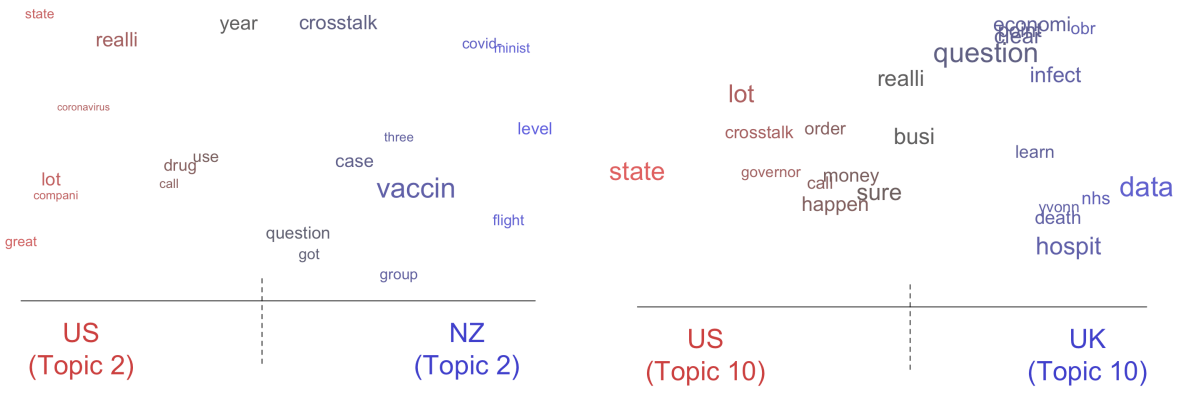


Table 10: Comparison between countries for topics 2 (left) and 10 (right)

Furthermore, the overall distribution of emotion terms at a country-level followed a very similar pattern, as shown in the bar plot for the US below.

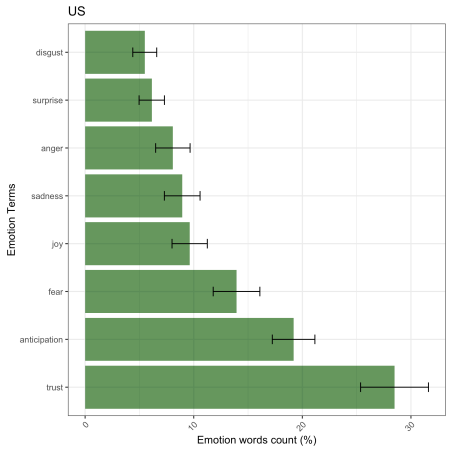


Table 11: Comparison between proportions of emotion terms used in the US corpus

Trust and anticipation were overwhelmingly present in all countries’ speeches, followed by fear. The next two topics were joy and sadness, whose order varied between countries, but their proportions were very close to each other in all cases[35]. This distribution is in line with what would be expected in response to a global pandemic. Fear and anticipation of the outcome are strong, but

[35] None of the countries showed a relatively significant amount of anger, surprise, or disgust terms.

leaders project trust in order to demonstrate control over the situation. Joy and sadness are both present as the situation betters or worsens respectively.

More interesting results can be seen in the plots of the distribution over time, particularly in the case of the UK, shown below.

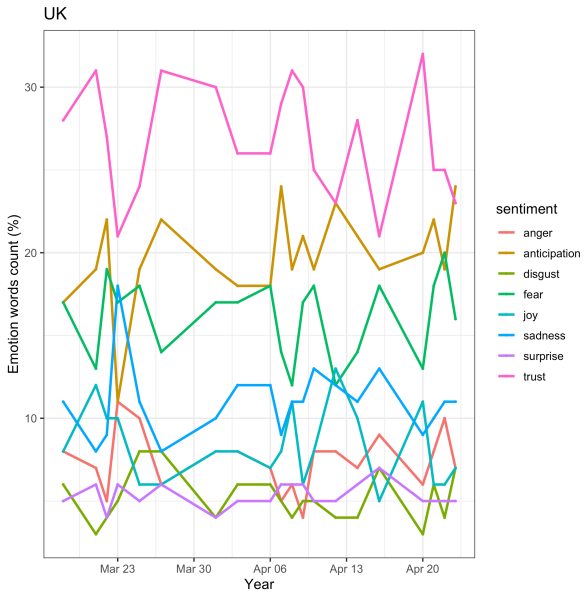


Table 12: Proportion of emotion words used over time in UK transcripts

On March 23, there was a sharp dip in trust and anticipation, with an accompanying increase in sadness and anger. The title of the article on this day is "*Boris Johnson Coronavirus Speech Transcript – Announces UK Lockdown: "You Must Stay at Home*". The plot reflects the change in emotion that occurs with the introduction of sudden, strict measures. A peak in anticipation and fear can also be seen just prior to this, on March 22nd, when Johnson warned that *"Italy-style lockdown is possible"*. Additionally, there is less variation in the proportion between emotions between March 27th and April 6th. On March 27th, Boris Johnson announced that he had coronavirus, and the press conferences following that date were held by UK health officials. This could reflect the difference in the way a country leader addresses the people of that country,

as opposed to the way a healthcare professional initially delivers national updates on a medical situation[36].

Discussion

The results overall confirmed our hypothesis that there are differences in country leaders' approaches to controlling a pandemic, and in how they portray themselves and the situation in extreme times. The countries whose transcript topics focused on more specific and concrete aspects of the pandemic, such as quantifying the number of cases and outlining health resources, seemed to mitigate the spread of the virus more efficiently. Additionally, changes in emotion could indicate political response adjustments, as evidenced in the emotion words analysis above.

However, there are caveats to this analysis. First, our data corpus was relatively small, especially small within countries. Countries such as New Zealand and Australia had ten or less instances in the corpus, while nearly half of our corpus (47.5%) was comprised of US documents. This imbalance could be indicative of increased press conferences held by the US, or it could also be a shortfall of the data source, which may have more access to US press conferences for transcription. Additionally, these press conferences all represent a specific extreme circumstance affecting the world. While this does make sense in the context of our question of analyzing emergency response among countries, these results cannot directly be extended to reflect each country leader's general policies and practices. The speeches given and steps taken are done so in strenuous, unforeseen circumstances, which likely affect the response.

Future Work

Further research should be done to incorporate more data at the given time range, as well as newer data to analyze how the response changed as conditions improved. This approach can also be extended to other countries and other text sources, as data becomes available. Another

[36] The subsequent change in the pattern could be due the officials' adapting to the press conferences. Similar plots for the remaining countries are summarized in Table A in the Appendix.

aspect of this research question that we wanted to explore was document readability. Did the Flesch score of each transcript affect the reception of the political piece? And if so, how did the public perceive it? Additionally, we wanted to explore the role that public sentiment played in terms of the topics that leaders discussed, and the potential effect that public sentiment has on delivery. Finally, we hope to focus our future efforts on finding a way to accurately and holistically measure country-wide policies, and relate this metric to our work. Do specific topics in political coronavirus briefings have any measurable impact on country policies? And if so, what do these effects tell us about how our leaders can better manage public health crises in the future?

References

Blei, David M., et al. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, Jan. 2003, pp. 993–1022., http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Denny, Matthew J., and Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis*, vol. 26, no. 2, 2018, pp. 168–189., doi:10.1017/pan.2017.44.

Dobrovolskyi, Hennadii, and Nataliya Keberle. "Principal Component Analysis in Topic Modelling of Short Text Document Collections." *CEUR Workshop Proceedings*, http://ceur-ws.org/Vol-1851/paper-8.pdf.

Dong, Mengying, et al. "Understand Research Hotspots Surrounding COVID-19 and Other Coronavirus Infections Using Topic Modeling." 2020, doi:10.1101/2020.03.26.20044164.

Kao, Anne. "Latent Semantic Analysis and Beyond." *Handbook of Research on Text and Web Mining Technologies*, pp. 546–570., doi:10.4018/978-1-59904-990-8.ch032.

Elliott, Thomas. *Topic Modelling*. 23 Jan. 2016, thomaselliott.me/pdfs/earl/topic_modeling.html.

Gebeyaw, Mesfin. "*Parsing Text for Emotion Terms: Analysis & Visualization Using R.*" DataScience , 13 May 2017, datascienceplus.com/parsing-text-for-emotion-terms-analysis-visualization-using-r/.

Kabir, Yasin, and Sanjay Madria. "*CoronaVis: A Real-Time COVID-19 Tweets Analyzer.*" 29 Apr. 2020, doi:arXiv:2004.13932v1 .

Liu, Qian, et al. "Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach." *National Center for Biotechnology Information*, U.S. National Library of Medicine, 18 Apr. 2020, www.ncbi.nlm.nih.gov/research/coronavirus/publication/32302966.

"Principal Component Analysis for Special Types of Data." *Principal Component Analysis Springer Series in Statistics*, pp. 338–372., doi:10.1007/0-387-22440-8_13.

Roberts, Margaret E., et al. "Stm: An R Package for Structural Topic Models." *Journal of Statistical Software*, vol. 91, no. 2, 2019, doi:10.18637/jss.v091.i02.

Saif M. Mohammad and Peter Turney. (2013), "*Crowdsourcing a Word-Emotion Association Lexicon.*" Computational Intelligence, 29(3): 436-465. doi: 10.1111/j.1467-8640.2012.00460.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x

Appendix

Appendix A: Principal Component Analysis and Dimension Reduction

In addition to LDA and STM, we also ran principal component analysis (PCA) on the trimmed DFM. Generally, topics that are discovered using probabilistic modeling and PCA can be used to represent shorter documents as vectors of real numbers, and the associated terms can then be used to "search for new documents" that extend each collection[37]. However, the principal components that result from this analysis are often difficult to interpret, and may not always be useful for longer texts, such as transcripts[38]. Compared to LDA, which may be described as "discrete PCA", pure PCA often shows results that are less intuitive, and occasionally lower performing,[39] as seen in the table below.

	PC1	PC2	PC3	PC4	PC5
thank	0.0201939405	0.0007608292	-0.0044494429	0.0089035439	-0.0015390189
much	0.0290046605	0.0133419926	-0.0040063289	-0.0001040891	0.0017718146
later	0.0195781111	0.0109261595	-0.0055975471	-0.0107757035	0.0075208260
even	0.0377355759	0.0145801504	-0.0061343551	-0.0004495281	-0.0005723181
expect	0.0254319211	-0.0044173116	0.0059805358	-0.0077608623	-0.0001538596
hous	0.0144205249	-0.0161386493	-0.0007654387	-0.0151760826	0.0048423024
pass	0.0009555819	-0.0011202359	0.0101919883	-0.0049571408	-0.0096259279
paycheck	0.0142061971	-0.0197528184	-0.0177947953	-0.0007118259	0.0018932901
protect	0.0226019776	-0.0279371278	0.0113291190	0.0133086644	-0.0004189988
program	0.0211438121	-0.0227715970	-0.0223224099	-0.0096225270	-0.0158387668

Table 6: PCA Topic-Term Decomposition and Factor Analysis

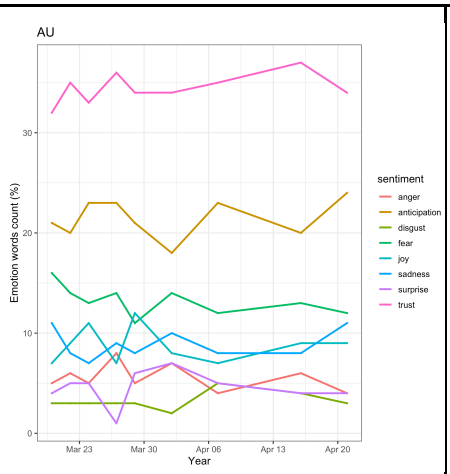
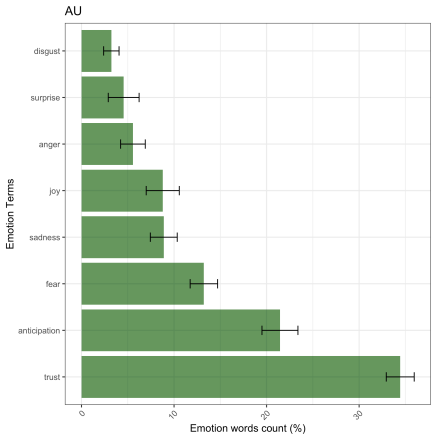
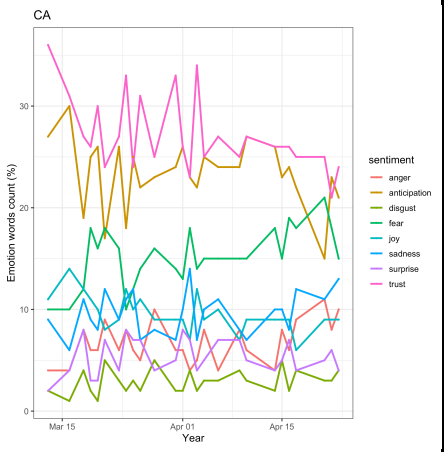
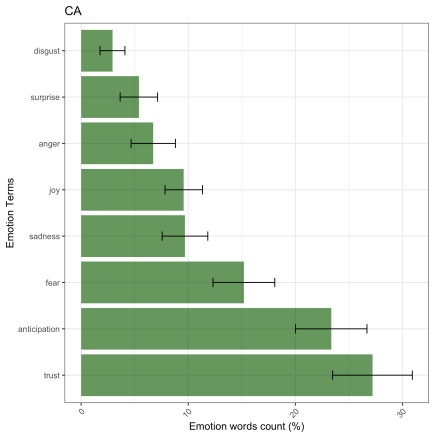
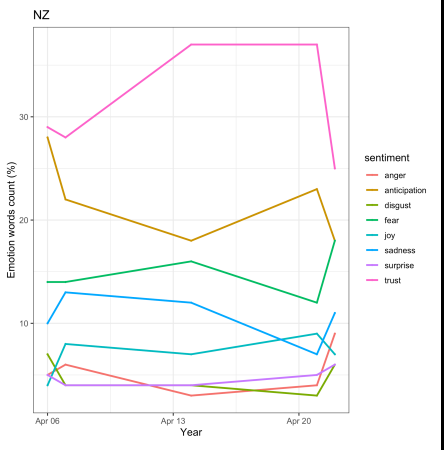
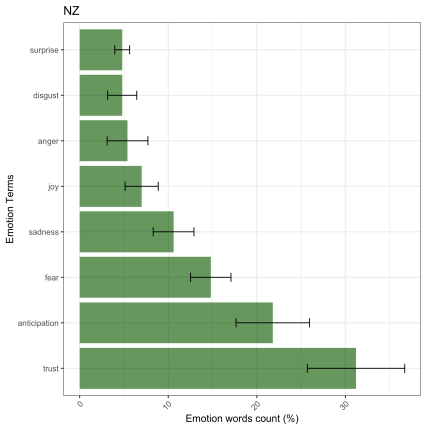
Appendix B: Emotion Visualizations per Country

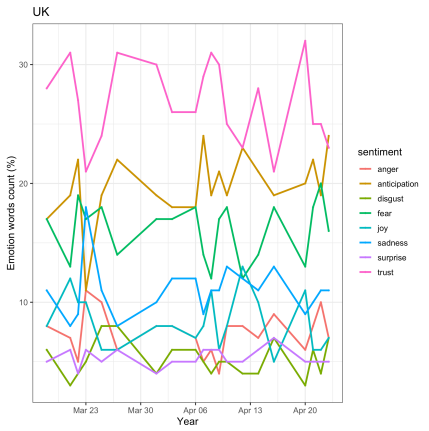
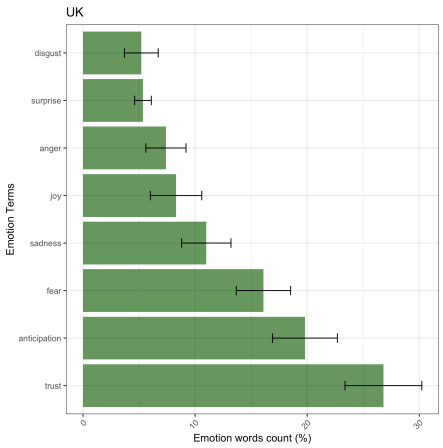
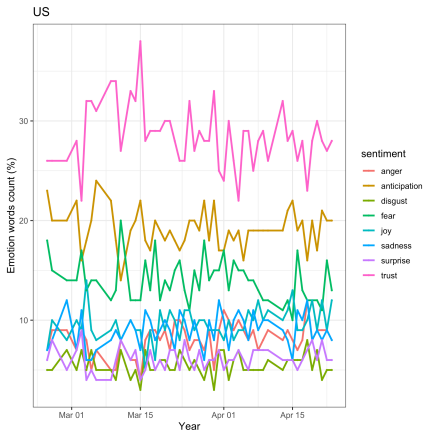
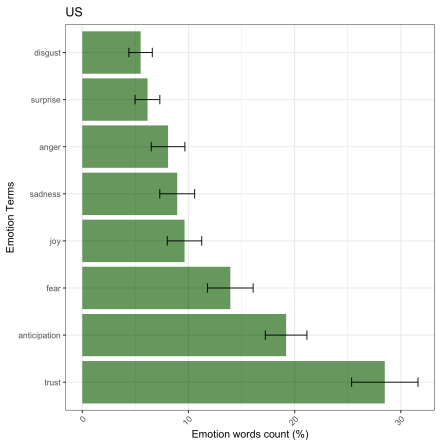
Country	Distribution over Time	Distribution within Country

[37] http://ceur-ws.org/Vol-1851/paper-8.pdf

[38] https://m-clark.github.io/sem/topic-models.html

[39] https://m-clark.github.io/sem/topic-models.html

Australia		
Canada		
New Zealand		

United Kingdom		
United States		

Appendix C: LSA “Coronavirus” Associated Words

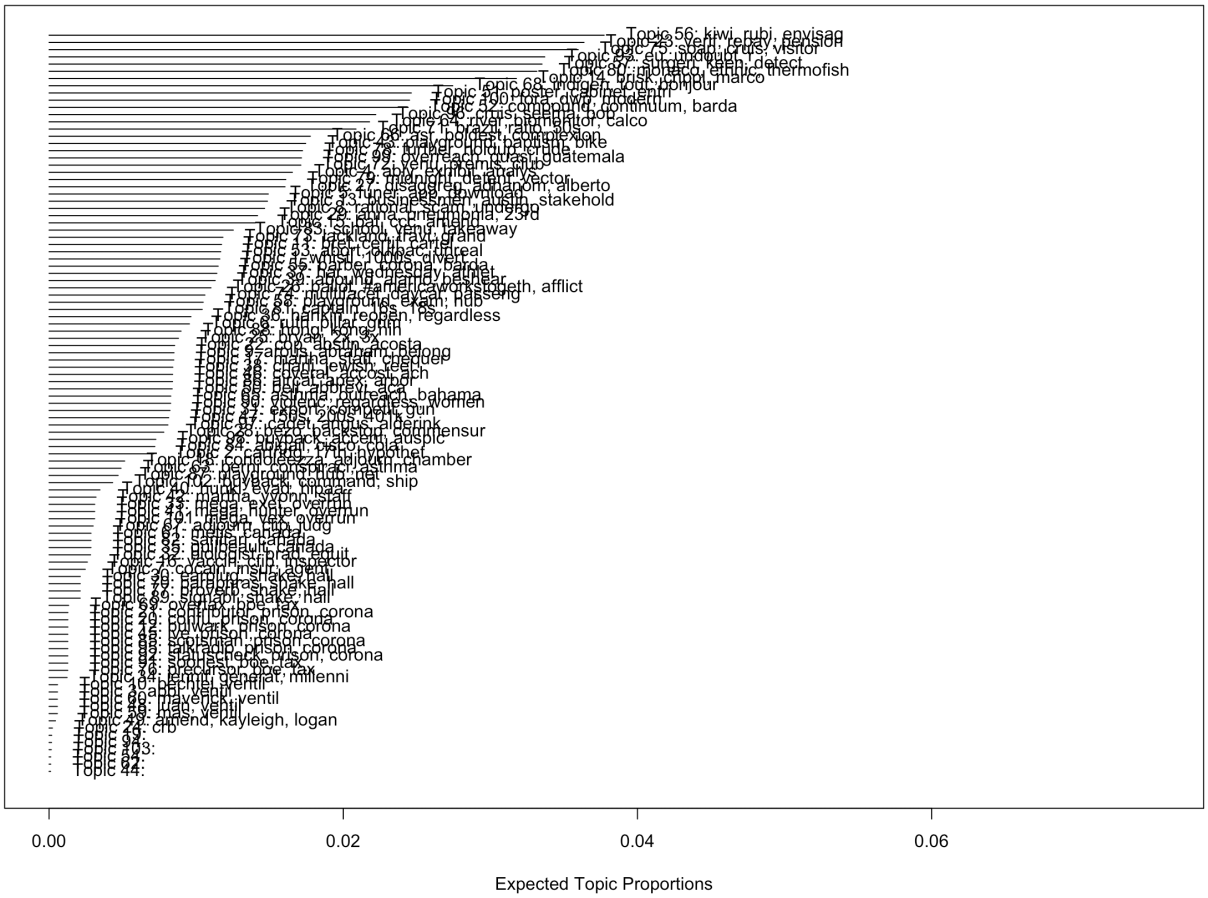
nhs	work	come	countri	thank	can	just	home	want	peopl
0.9820540	0.9777941	0.9775221	0.9768404	0.9753108	0.9736450	0.9718401	0.9694531	0.9687810	0.9683781

Table 1: Top 10 Closest Words to “coronavirus” in UK Subset

ethnic	intervent	organ	phone	rough	excel	context	train	foot	spoke
0.9927258	0.9908264	0.9768600	0.9589660	0.9589660	0.9563788	0.9563788	0.9563788	0.9563788	0.9563788

Table 2: Top 10 Closest Words to “coronavirus” in NZ Subset

Appendix D: STM Top Topics and Their Relative Proportions


Top Topics
0.00
0.02
0.04
0.06
Expected Topic Proportions