2020

Predicting Performance of Online Media Content

SPINGBOARD CAPSTONE II

STUDENT LIUDMILA KHALITOVA, SUPERVISOR VARUN BHATIA

FINAL REPORT

Public relations practitioners have always been looking for ways to increase positive media attention to their organizations. One common practice includes media pitching and distribution of a company's news to journalists and media outlets via press releases. While public relations managers get to decide what kind of news their company will be sharing with media, journalists and editors are the gatekeepers deciding whether the audience of their outlets will actually see that news. Relevance and audience interest in a particular topic are one of the common factors that media professionals take into account when making editorial decisions. If a corporation's public relations department knows exactly what kind of their corporate news generate the most interest among media audiences, they would be able to pitch outlets more successfully by addressing the topics that attract readers the most.

I work for Cision, a media research company that helps public relations practitioners track and assess their companies' visibility in traditional outlets and social media. By using the data collected for one of its clients, Service Now, I created a tool to help other clients better understand their media coverage and predict what kind of topics related to their company would perform better than others in terms of attracting page views and unique visitors. In addition, the tool helps assess topics in terms of their lifespan, that is, what topics remain relevant to news consumers for a longer period.

Specific uses across projects within Cision Insights:

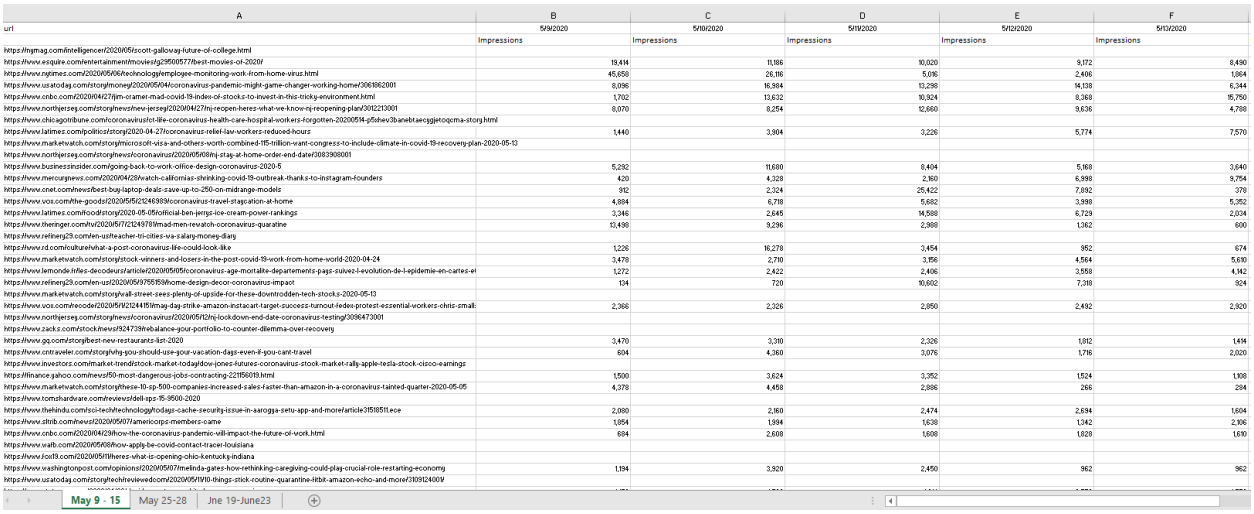
1. Topic identification. Cision's software has the ability to determine the topic of news media content; however, it is based on pre-defined keywords and search strings, which, in turn, requires topics to be pre-defined by clients. My project offers a data-driven approach to topic identification which can help clients to better understand the context in which their organization is being covered by news media and guide the initial topic selection for further tracking.
2. Predicting the performance of news media content after its publication. The project determines half-life of a news media article and whether it depends on its topic. Knowing how long it takes for an article to reach half of its views or unique users can be useful in two major ways. First, public relations practitioners and/or analysts can quicker assess and predict overall campaign performance, without having to wait till the campaign is over. Second, it can reduce the system load of the Cision Impact platform by limiting the tracking period to articles' average half-life.

Data Wrangling

1. The data set.

The original data set represented an Excel file exported from the Cision Impact platform. The file had three sheets with date ranges: May 9-15, 2020, May 25-28, and June 19 through 23 of

2020. Each sheet had a column with URLs, columns with dates and sub-columns with metrics, Impressions (how many times an article was viewed) and Unique Users (how many unique users have visited the page) on each day.



	A	B	C	D	E	F
1	url	5/9/2020	5/10/2020	5/11/2020	5/12/2020	5/13/2020
2		Impressions	Impressions	Impressions	Impressions	Impressions
3	https://ngmag.com/intelligencer/2020/05/scott-galloway-future-of-college.html					
4	https://www.esquire.com/entertainment/movies/g29500577/best-movies-of-2020/	19,414	11,186	10,020	9,172	8,490
5	https://www.nytimes.com/2020/05/06/technology/employee-monitoring-work-from-home-virus.html	45,658	26,116	5,016	2,406	1,864
6	https://www.usatoday.com/story/money/2020/05/04/coronavirus-pandemic-might-game-changer-working-home/3061862001	8,096	16,984	13,286	14,138	6,344
7	https://www.cnbc.com/2020/04/27/jim-cramer-mad-covid-19-index-of-stocks-to-invest-in-this-tricky-environment.html	1,702	13,632	10,924	8,368	15,750
8	https://www.northjersey.com/story/news/new-jersey/2020/04/27/nj-reopen-heres-what-we-know-nj-reopening-plan/3012213001	8,070	8,254	12,660	9,636	4,788
9	https://www.chicagotribune.com/coronavirus/ct-life-coronavirus-health-care-hospital-workers-forgotten-20200514-p5shev3banebtaecggjjetoqcma-story.html					
10	https://www.latimes.com/politics/story/2020-04-27/coronavirus-relief-law-workers-reduced-hours	1,440	3,904	3,226	5,774	7,570
11	https://www.marketwatch.com/story/microsoft-visa-and-others-worth-combined-115-trillion-want-congress-to-include-climate-in-covid-19-recovery-plan-2020-05-13					
12	https://www.northjersey.com/story/news/coronavirus/2020/05/06/nj-stay-at-home-order-end-date/3083906001					
13	https://www.businessinsider.com/going-back-to-work-office-design-coronavirus-2020-5	5,292	11,680	8,404	5,168	3,640
14	https://www.mercurynews.com/2020/04/29/watch-californias-shrinking-covid-19-outbreak-thanks-to-instagram-founders	420	4,328	2,160	6,998	9,754
15	https://www.cnet.com/news/best-buy-laptop-deals-save-up-to-250-on-midrange-models	912	2,324	25,422	7,892	378
16	https://www.vox.com/the-goods/2020/5/5/21246989/coronavirus-travel-staycation-at-home	4,884	6,718	5,682	3,998	5,352
17	https://www.latimes.com/food/story/2020-05-05/official-ben-jerrys-ice-cream-power-rankings	3,346	2,645	14,588	6,729	2,034
18	https://www.theringer.com/tv/2020/5/7/21249769/mad-men-revatch-coronavirus-quarantine	13,498	9,296	2,988	1,362	600
19	https://www.refinery29.com/en-us/the-bachelor-tri-cities-wa-salary-money-diary					
20	https://www.rd.com/culture/what-a-post-coronavirus-life-could-look-like	1,226	16,278	3,454	952	674
21	https://www.marketwatch.com/story/stock-winners-and-losers-in-the-post-covid-19-work-from-home-world-2020-04-24	3,478	2,710	3,156	4,564	5,610
22	https://www.lemonde.fr/les-decodeurs/article/2020/05/05/coronavirus-age-mortalite-departements-pays-suivez-l-evolution-de-l-epidemie-en-cartes-et	1,272	2,422	2,406	3,558	4,142
23	https://www.refinery29.com/en-us/2020/05/9755159/home-design-decor-coronavirus-impact	134	720	10,602	7,318	924
24	https://www.marketwatch.com/story/wall-street-sees-plenty-of-upside-for-these-downtrodden-tech-stocks-2020-05-13					
25	https://www.vox.com/recode/2020/5/9/21244151/may-day-strike-amazon-instacart-target-staceess-turnout-fedex-protest-essential-workers-chris-smalls	2,366	2,326	2,850	2,492	2,920
26	https://www.northjersey.com/story/news/coronavirus/2020/05/12/nj-lockdown-end-date-coronavirus-testing/3096473001					
27	https://www.zacks.com/stock/news/924739/rebalance-your-portfolio-to-counter-dilemma-over-recovery					
28	https://www.gq.com/story/best-new-restaurants-list-2020	3,470	3,310	2,326	1,912	1,414
29	https://www.cntraveler.com/story/why-you-should-use-your-vacation-days-even-if-you-cant-travel	604	4,360	3,076	1,716	2,020
30	https://www.investors.com/market-trend/stock-market-today/dow-jones-futures-coronavirus-stock-market-rally-apple-tesla-stock-cisco-earnings					
31	https://finance.yahoo.com/news/50-most-dangerous-jobs-contracting-221156019.html	1,500	3,624	3,352	1,524	1,108
32	https://www.marketwatch.com/story/these-10-sp-500-companies-increased-sales-faster-than-amazon-in-a-coronavirus-tainted-quarter-2020-05-05	4,378	4,458	2,886	266	284
33	https://www.tomshardware.com/reviews/dell-xps-15-9500-2020					
34	https://www.thehindu.com/sci-tech/technology/todays-cache-security-issue-in-aarogya-setu-app-and-more/article31518511.ece	2,080	2,160	2,474	2,694	1,604
35	https://www.sltrib.com/news/2020/05/07/americorps-members-came	1,854	1,994	1,638	1,342	2,106
36	https://www.cnbc.com/2020/04/29/how-the-coronavirus-pandemic-will-impact-the-future-of-work.html	684	2,608	1,608	1,828	1,610
37	https://www.wafb.com/2020/05/08/how-apply-be-covid-contact-tracer-louisiana					
38	https://www.fox19.com/2020/05/11/theres-what-is-opening-ohio-kentucky-indiana					
39	https://www.washingtonpost.com/opinions/2020/05/07/melinda-gates-how-rethinking-caregiving-could-play-crucial-role-restarting-economy	1,194	3,920	2,450	962	962
40	https://www.usatoday.com/story/tech/reviewed/com/2020/05/11/10-things-stick-routine-quarantine-fitbit-amazon-echo-and-more/3109124001/					

May 9 - 15 | May 25-28 | Jne 19-June23

2. Data preparation

2.1. Merging and cleaning. The first step was to merge data from each sheet into one data frame by placing all URLs into one column while preserving each URL's information about daily views and unique visitors. If no data was available for a certain article on a certain date, NaNs were replaced by zeros. In addition, I identified duplicate URLs (articles that continued to generate views longer than others and appeared in multiple sheets). Their Impressions and Reach data was moved into one row, and the duplicate rows were removed. The cleaned and merged dataset had 25,440 rows and 33 columns.

2.2. URL parsing. Because full text of articles was not available, I relied on the URL information to determine what the article was about. News outlets typically include either the headline of the article or its key words into the article URL address. For example, a New York Times article titled "How My Boss Monitors Me While I Work From Home" discussed how companies use software to monitor their employees working from home during the coronavirus pandemic. The article had the following URL: https://www.nytimes.com/2020/05/06/technology/employee-monitoring-work-from-home-virus.html

From the URL, it is clear that the article was about the monitoring technologies, employees and work from home. I used Urllib's Url Parse module which splits a URL into the following components: scheme (or protocol, such as http or https), netloc (network location path, or domain name, for example, www.nytimes.com), path (hierarchical path such as "2020/05/06/technology/employee-monitoring-work-from-home-virus.html". The [path] column contained the keyword information I was interested in.

2.3. Keyword preparation. I preprocessed the [path] column by deleting special characters such as /, ?. digits, and tags, and removed stop words common in the English language with NLTK (Natural Language Toolkit). Next, I identified the most important keywords within each URL by using the TF-IDF keyword extraction method via Python's Scikit-learn package. In essence, the method identifies the words that are unique for a specific document (in our case, a URL path) while ignoring the words appearing in 85% of the documents. I limited the number of unique words to be extracted from each URL to 5. This yielded a column of tuples containing keywords with their respective TF-IDF values indicative of their importance in the document:

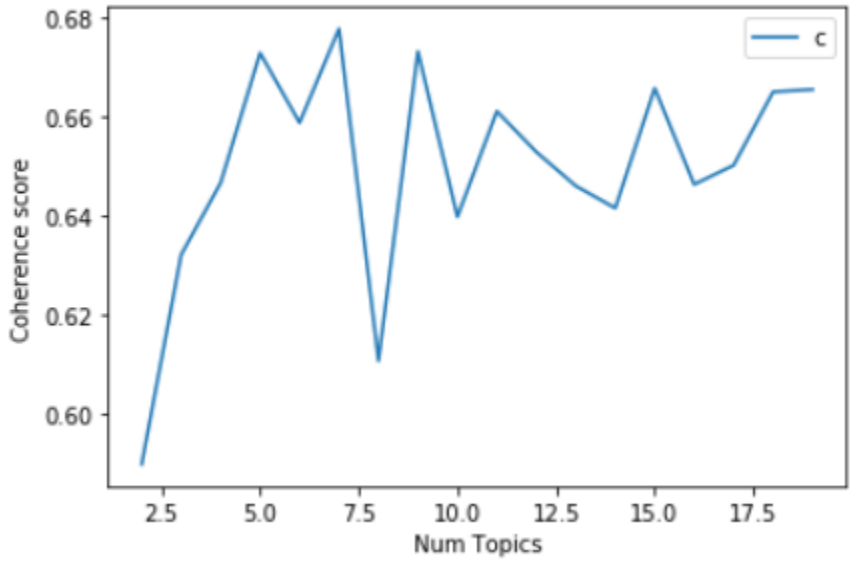
doc	keywords
intelligencer scott galloway future of college html	{'intelligencer': 0.54, 'galloway': 0.493, 'scott': 0.447, 'college': 0.404, 'future': 0.32}
entertainment movies g best movies of	{'movies': 0.892, 'entertainment': 0.332, 'best': 0.308}
technology employee monitoring work from home viru	{'monitoring': 0.529, 'employee': 0.486, 'virus': 0.358, 'technology': 0.352, 'work': 0.341}
story money coronavirus pandemic might game change	{'changer': 0.527, 'game': 0.452, 'money': 0.345, 'working': 0.311, 'pandemic': 0.306}
jim cramer mad covid index of stocks to invest in this tr	{'tricky': 0.446, 'mad': 0.387, 'environment': 0.366, 'cramer': 0.366, 'jim': 0.33}
story news new jersey nj reopen heres what we know r	{'nj': 0.69, 'jersey': 0.363, 'reopening': 0.256, 'reopen': 0.256, 'know': 0.256}
coronavirus ct life coronavirus health care hospital worl	{'xhev': 0.393, 'banebtaecygjetoqcma': 0.393, 'forgotten': 0.381, 'coronavirus': 0.338, 'hospital': 0.316}
politics story coronavirus relief law workers reduced hc	{'reduced': 0.468, 'relief': 0.426, 'law': 0.414, 'hours': 0.386, 'politics': 0.31}
story microsoft visa and others worth combined trillion	{'combined': 0.385, 'include': 0.359, 'trillion': 0.339, 'congress': 0.339, 'climate': 0.31}
story news coronavirus nj stay at home order end date	{'nj': 0.469, 'date': 0.422, 'order': 0.385, 'end': 0.367, 'stay': 0.356}
going back to work office design coronavirus	{'going': 0.511, 'design': 0.483, 'office': 0.394, 'back': 0.39, 'work': 0.336}
watch californias shrinking covid outbreak thanks to ins	{'shrinking': 0.419, 'californias': 0.419, 'founders': 0.402, 'thanks': 0.376, 'instagram': 0.372}
news best buy laptop deals save up to on midrange mo	{'midrange': 0.496, 'models': 0.407, 'laptop': 0.389, 'save': 0.369, 'deals': 0.365}

3. Grouping keywords into topics.

The next step was to group the identified keywords into broader topics.

3.1. Preprocessing. I repeated the preprocessing steps for the [keywords] column to get rid of digits and special characters and lemmatized the cleaned keywords with NLTK Tokenize package.

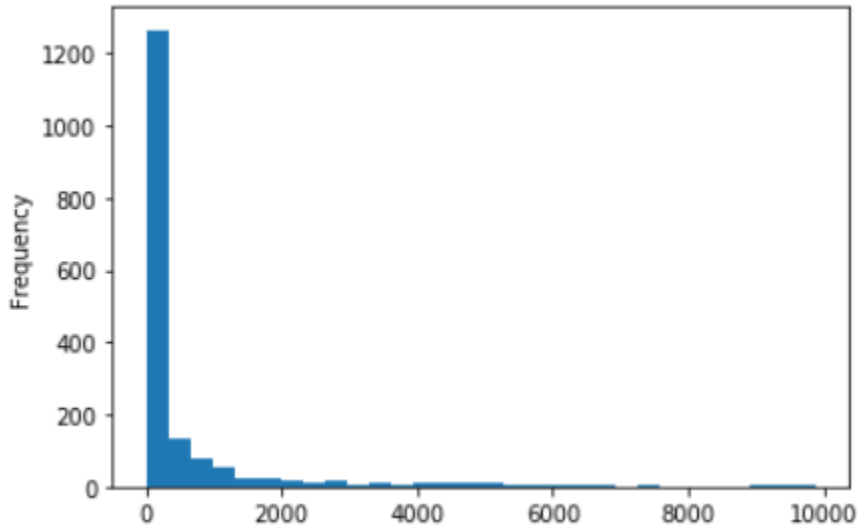
3.2. Topic modeling and model selection. I built multiple LDA models using Genism to identify the optimal number of topics based on their coherence scores.



I selected a 6-topic model, and based on their keyword composition, labeled these topics as follows: Business and digital technology, COVID-19 updates, Employee wellness, Remote work, Sports and competition, and Stock market performance. For further statistical analyses, I selected each topic's most representative articles based on their topic's percent contribution being above 60%.

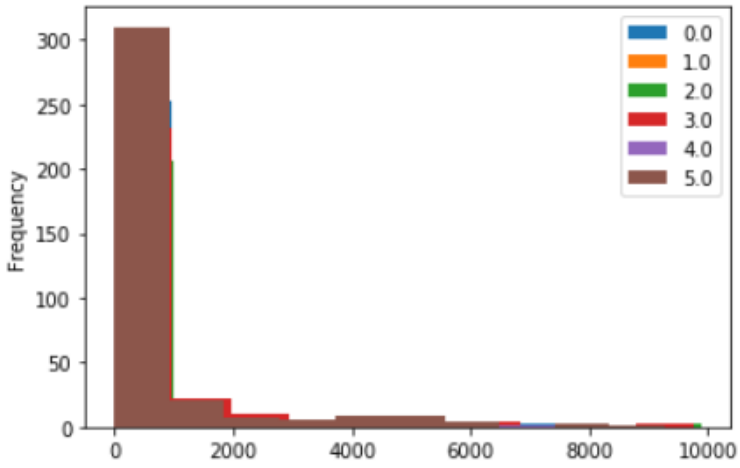
Exploratory Data Analysis

1. Data cleaning. For each article, I calculated the sum of daily impressions and visitors and dropped the ones that had none. The dataset resulted in 1754 articles.
2. Data Exploration. I created a histogram of total impressions distributions to see if the values were normally distributed. I cut off some of the obvious outliers with values exceeding 10,000 impressions but the data was still highly skewed to the right.

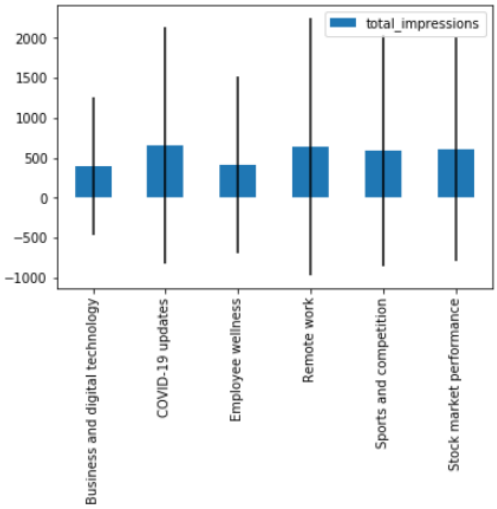


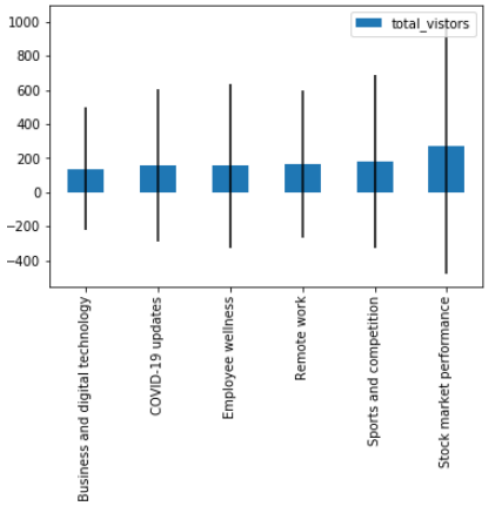
I further explored the frequency of each topic's occurrence to make sure topics are distributed more or less equally in the dataset. It turned out to be the case, topic frequencies ranged from to 241 to 365 articles per topic. Moreover, the shapes of impression distributions were almost identical across topics:

```
import matplotlib.pyplot as plt
df.groupby(['Dominant_Topic'])['total_impressions'].plot(kind='hist', bins=10)
plt.legend()
plt.show()
```



I plotted each topic's mean impressions and unique visitors counts with error bars to see if there are any visible differences across topics:





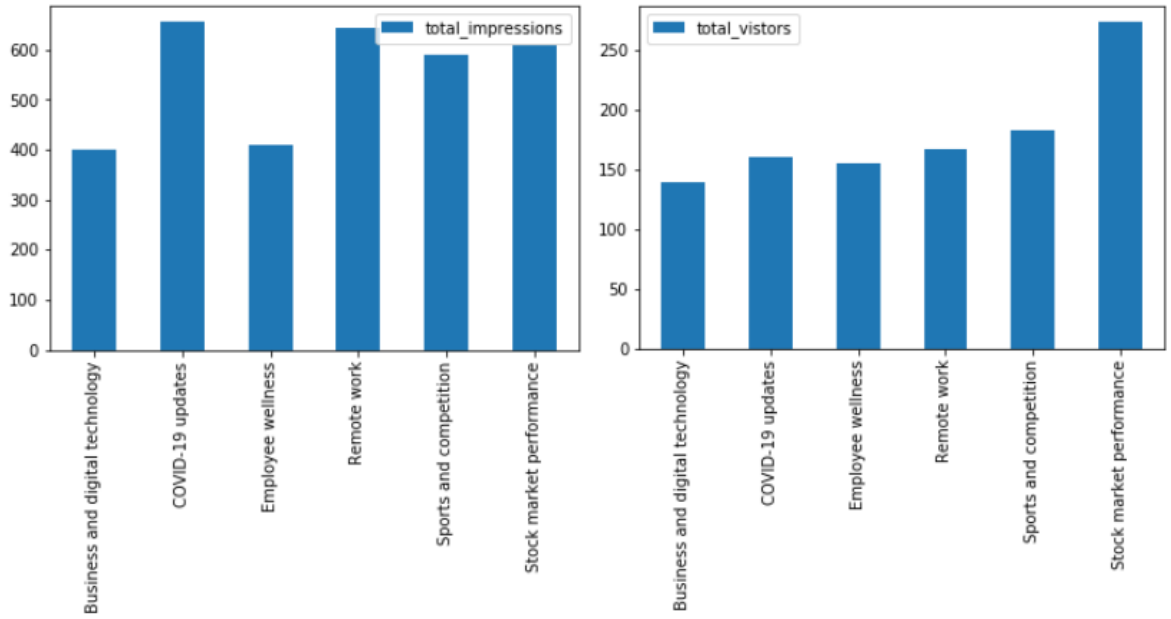
Wide confidence intervals further suggest that the data is not distributed normally which needs to be taken into account when performing statistical tests. To correct for this, I chose to perform bootstrapping on 1,000 samples with a 95% bias-corrected confidence interval.

Further Analysis

1. Do some topics attract audience attention better than others?

   The charts below show each topic's average numbers of impressions and unique visitors per article. The impressions chart suggests that articles in the Business and Technology and

Employee Wellness categories, on average, receive 200 fewer views than articles in other categories. The Unique Visitors chart suggests that articles in the Stock Market category, on average, attract 100 more users than articles on other topics.



I want to know if these differences are indeed due to the differences in topics and whether they are statistically meaningful.

1.1 Taking into account alternative explanations.

The data was collected from more that three hundred different online outlets ranging from the New York Times, Forbes, and Bloomberg to the niche professional outlets such as Recruiter Box. Therefore, it is entirely possible that the number of views and visitors has less to do with the article's content per se and more to do with the outlet's scope and, hence, its reach. This means that in order to parse out the effects of topic on article's views and the number of unique visitors, outlet's reach needs to be controlled for.

1.2 Data preparation.

To control for outlet's reach I used Cision's proprietary Horreum platform to get a list of online outlets in English with their average daily views. Then I added daily views for outlets in my dataset by matching its [netloc] column with the [domain] column of the Horreum dataset.

1.3. Statistical analyses.

To see if differences in user and impression count across topics were meaningful, while controlling for the effects of outlets' reach, I performed an analysis of covariance (ANCOVA) on my dataset using SPSS Statistics 25. To correct for the skewed distribution, I performed bootstrapping on a 1,000 samples with 95% bias-corrected confidence intervals. Topic

(categorical variable) was included into the model as a factor while outlet's reach was entered as a covariate.

1.4. Results

Effects of Topic on article's impressions. Bootstrapped test for parameter estimates indicated that only Business and Technology significantly differed from other topics in terms of their ability to generate impressions. This topic, on average, received 209 views fewer than other topics (p = .028). Pairwise comparisons revealed that in comparison with Remote Work, Business and Technology received 244 fewer views (p = .043); in comparison with COVID-19 Updates it had 255 fewer views (p = .015) and 209 fewer views than Stock Market Performance (p = .028). There were no statistically significant differences between Business and Technology and Sports and Competition and Business and Technology and Employee Wellness.

Topic		Mean Difference	Sig (2-tailed)	95% CI upper	95% CI lower
Business and Technology	Sports and Competition	-184.939	.067	-392.080	4.187
	Remote Work	-244.120	**.043**	-506.284	-34.702
	COVID-19 Updates	-254.796	**.015**	-456.280	-59.432
	Employee Wellness	-9.291	.922	-194.547	164.166
	Stock Market Performance	-208.699	**.028**	-393.359	-37.586

Effects of Topic on article's unique visitors. The ANCOVA test revealed that the effects of Topic on the number of unique visitors were more substantial that its effects on the number of views. More specifically, Stock Market Performance performed significantly better than every other topic. On average, it attracted 92 users more than Sports and Competition (p = .042), 133 users more than Business and Technology (p = .004), 106 users more than Remote Work (p = .033), 113 users more than COVID-19 (p = .018), and 118 users more than Employee Wellness (p = .022).

Topic		Mean Difference	Sig (2-tailed)	95% CI upper	95% CI lower
Stock Market Performance	Sports and Competition	91.651	**.042**	6.666	188.644
	Business and Technology	133.231	**.004**	47.445	226.766
	Remote Work	106.147	**.033**	14.864	209.245
	COVID-19 Updates	112.995	**.018**	19.662	210.155

Employee Wellness	118.123	**.022**	22.261	217.456

2. Do some topics retain audience attention better than others?

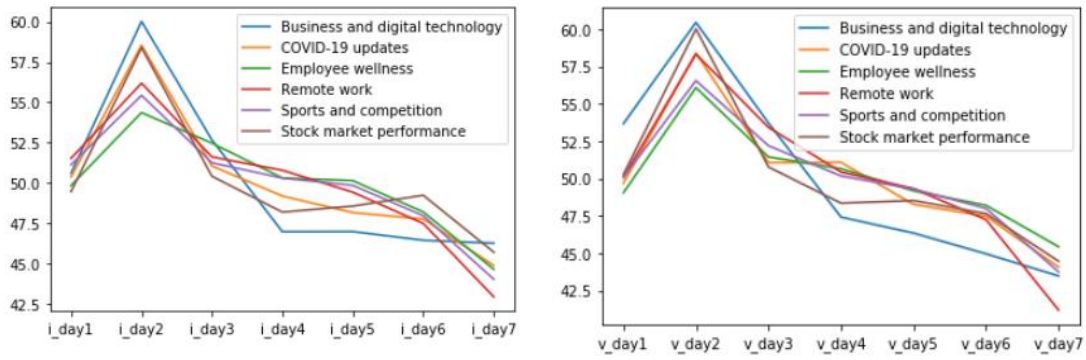
To answer this question, I explored each topic's half-life, that is, how long it takes for an article to reach half of its total views. The longer it takes, the longer the article's lifespan is, meaning that visitors continue to be interested in it despite the fact that it was published some time ago.

2.1 Data preparation.

For this analysis I used articles published between May 9 through 15 as this was the longest timeframe withing which the data was collected (7 days; the other two chunks had a 4-day interval). Some articles had thousands of more views than others; however, for this part of my study, I was interested in their ability to retain audience rather than initially attract it. Therefore, I converted each article's daily views and visitors into T-scores to make sure each article's views on the same day were measured on the same scale.

2.2. Data analysis.

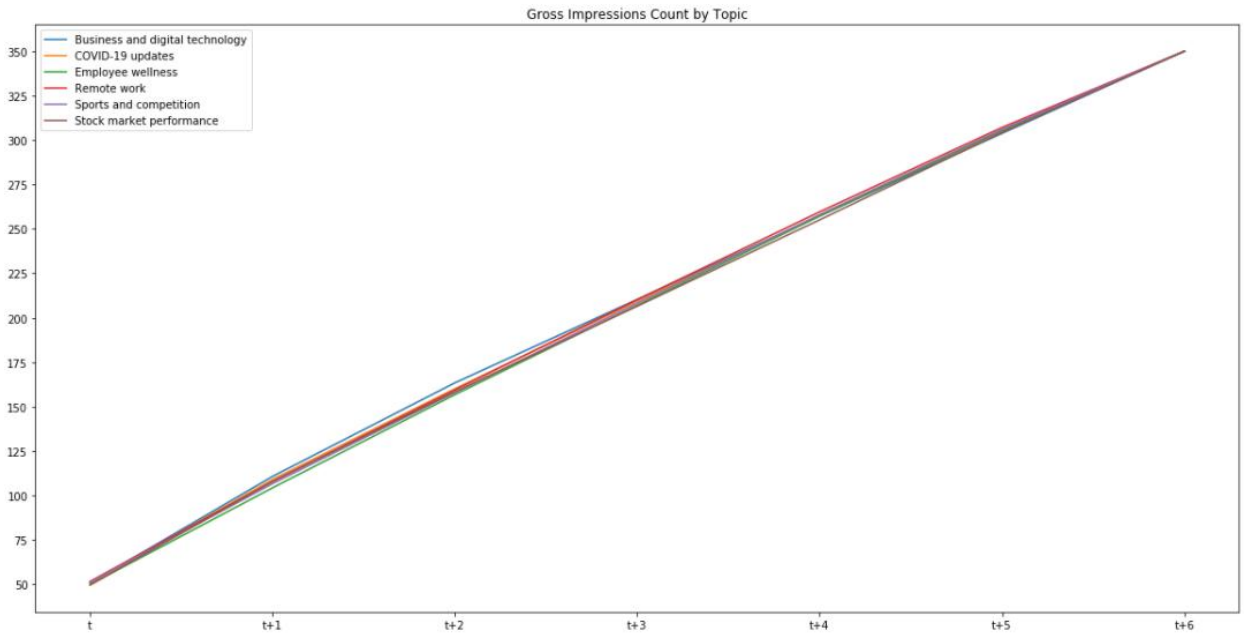
The charts below show daily impressions trends (left) and daily unique visitor trends (right) by topic.

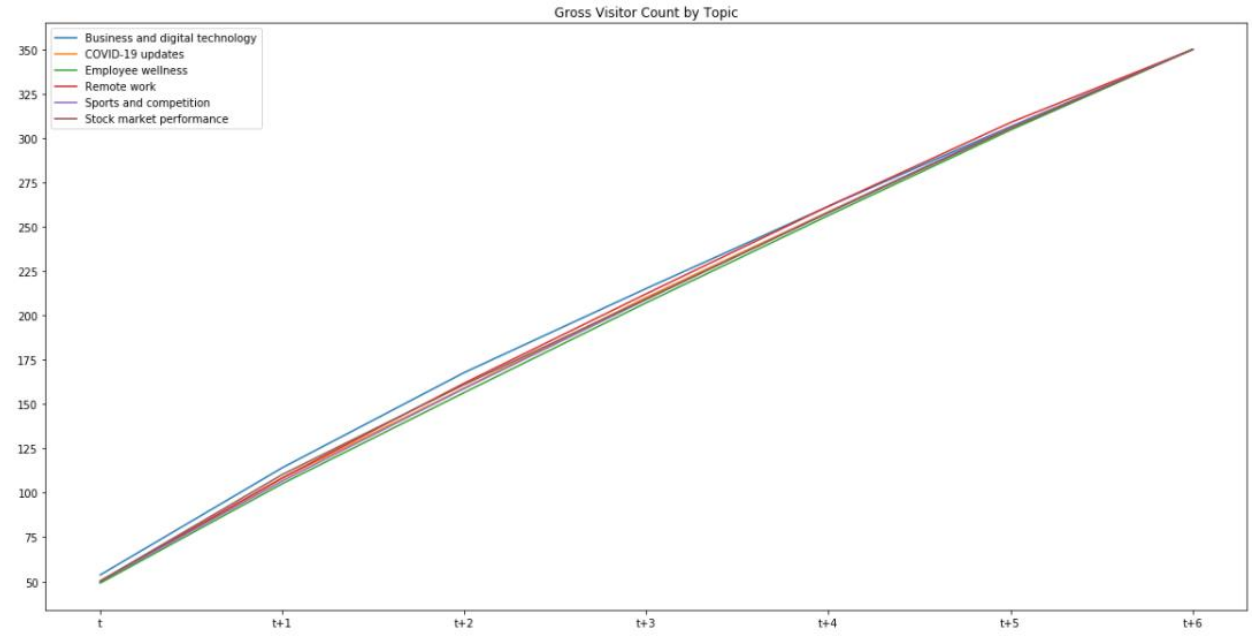


The charts suggest that both unique visitors and impressions across all topics peak around day 2, remain relatively stable on days 3 to 6, and drop from day 6. The decline for Business and Digital Technology seems to be the most dramatic. For each individual article, I calculated that article's half-lives in terms of both impressions and unique visitors by dividing their total impressions across all days by two (same for visitors). Both turned out to be equal to 175 across all topics. Then, I needed to plot gross or cumulative impressions and unique visitors to show their increase with each day and find out at what point in time each topic's cumulative impressions and visitor reach 175.

2.3 Results

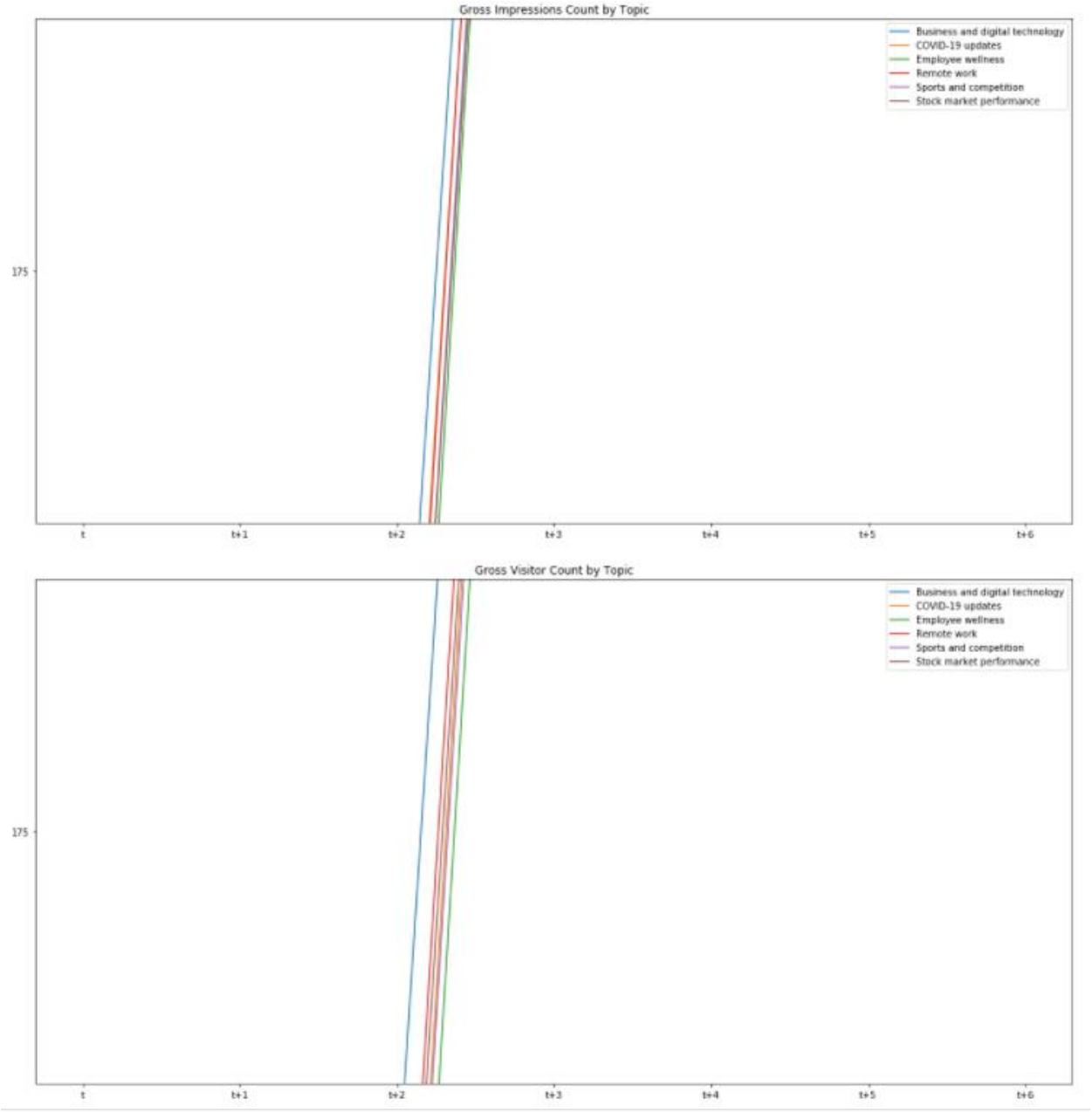
The plots below illustrate gross impressions (top) and gross unique visitors (bottom) count.





The charts suggest that regardless of topic, all articles reach their half-life between days 3 (t+2) and 4 (t+3) since their publication (t).

I have zoomed in the charts to see if there are any differences across topics:



The charts confirm that while Business and Technology does reach its half-life slightly quicker than other topics, still it remains within the same range as other topics (day 3 to 4).

Takeaways

- Among the topics extracted, Business and Technology news turned out to perform worse than most other topics in terms of both impressions and its lifespan.
- Stock Market Performance news performed better than all other topics in terms of its ability to attract unique visitors.
- Most articles receive half of their views and page visits on day 2 after the date of their publication.

- An article's topic did not significantly affect its lifespan, meaning that readers were loosing interest in online content at approximately the same rate across topics.

Future research

The dataset used in this study was limited to news media articles mentioning only one company, Service Now. Therefore, the generalizability of the findings presented across companies and industries is unknown. More research needed to confirm our findings regarding of the lifespan of online news content. In addition, I investigated the topics relevant specifically to Service Now, it is not clear how applicable they would be even within the same industry. Finally, some of the extracted topics are specific to the time frame when the data was collected (e.g. May – June 2020), when COVID-19 coverage was especially prominent. Therefore, the topic extraction procedure may need to be repeated for each client if the media landscape significantly changes.