Topic Evolution in College Newspapers Before and After COVId-19

Andrew Knox
Advisor: Dr. Natalia Khuri

August 10, 2020

1 Abstract

College news articles demonstrate how students react to the unfolding events happening in the country and around the world. Since the COVId-19 pandemic has forced dramatic change in the way people live, the main objective of this paper is to analyze the shift in news topics from before the pandemic to after. In this paper, newspaper topics are extracted using topic modeling algorithms, figures are produced to quantify the emergence and decline of topics per month, and sentiment analysis is run on each identified topic. Experimental results show that the pandemic diminished production of previously commonplace articles, and increased production of articles about societal change and racial injustice.

Contents

List of Figures

List of Tables

2 Introduction

The Covid-19 pandemic changed the way people live. Whether it was going to the store, seeing friends, or working from home, there were major differences in how individuals behaved. Students were adversely affected, and many had to return home for the end of the school year; however, student-led news organizations continued to publish articles. Examining the changes in written topics from student-led news organizations can help to better understand the impact that Covid-19 had on individuals and organizations.

Topics were discovered using topic modeling algorithms, a set of algorithms designed to extract words that are often grouped together. With these words, the viewer is able to identify a specific topic that the words comprise.

The main research objective is to analyze how college newspapers shifted in topics from pre-covid to post-covid.

Specific aim 1: To identify latent topics in over 1700 news articles in two time-slices: after covid-19, and before covid-19.

Specific aim 2: To analyze how prominent topics were, when they emerged and withdrew from the spotlight, and the sentiment that they were written with.

In order to achieve these aims, the author:

- gathered online articles from three different universities for a total of 1,782 articles.
- partitioned the dataset of articles into pre-covid and post-covid and ran Latent Dirichlet Allocation (LDA) on both datasets.
- visualized the percentage of articles that were produced each month for each topic.
- performed sentiment analysis on articles that comprised each topic.

3 Data and Methods

Articles are scraped from online using BeautifulSoup [Ric07] and stored on disk. From there, each article is pre-processed and a vocabulary list is generated from the most common words. Articles are transformed into a doc-word matrix and used for parameter tuning of the topic modeling algorithms. Then topic modeling algorithms are run with optimal parameters, and figures of words comprising each topic are produced. Afterwards, figures showing when each topic was most prevalent are produced, and sentiment analysis is run on the articles that comprise each topic.

All articles come from Wake Forest University, Stanford University, and Miami-Dade College, and were scraped from online sites wfuogb.com, stanforddaily.com, and mdcthereporter.com, respectively.

Dataset	No. Files	Avg. No. of Sentences	Avg. No. of Words
Wake Forest	655	37.97	4575.05
Stanford	719	35.84	4618.47
Miami-Dade	408	36.50	3622.15

Table 1: Basic information about each dataset

Articles were preprocessed using the NLTK module [LB02]. The specific steps were (1) making all letters lowercase, (2) removing all punctuation, (3) removing stop words using the NLTK stopwords module, and (4) lemmatization of the remaining words. Custom stop words specific to the dataset such as 'wake', 'forest', and 'stanford', were also removed. Additionally, words less than 3 letters long were considered to be stop words. Figure 1 shows the preprocessing steps and the experimental pipeline steps.

After preprocessing, documents were split and labeled either pre-covid or post-covid; articles written in February 2020 or later were considered post-covid, as the first news articles to contain the words 'COVId-19' and 'coronavirus' occurred in February. For both the pre-covid and post-covid datasets, a vocabulary was accumulated and all articles were transformed to a Term-Frequency matrix using Sklearn's Count Vectorizer[Ped+11]. The preprocessed articles were input into Sklearn's Latent Dirichlet Allocation algorithm and perplexity was used to test different parameters. After finding the best parameters, articles were grouped into topics, visualizations of the emergence of topics were created, and topics were analyzed using VADER sentiment analysis[Yan+19].

Perplexity was used to evaluate model performance. It is a statistical measure of how well a probability model predicts a sample. As applied to LDA, for different parameter values, the model is estimated. Then the theoretical word distributions represented by the topics are compared to the actual topic mixtures, or distribution of words in the documents. The lower the perplexity score, the better the model.

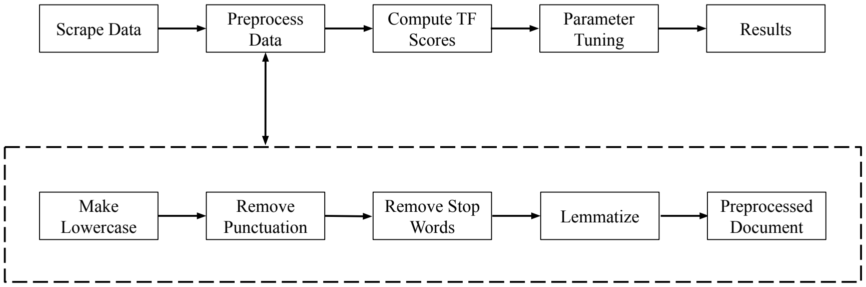

Figure 1: The experiment pipeline and preprocessing pipeline are shown.

The computational experiment to determine the best parameters for LDA modulated three parameters: n_components - the number of overall topics; doc_topic_prior (alpha) - the prior distribution that controls how many topics will be related to a single document; and topic_word_prior (beta) - the distribution that controls how many words will likely be related to a single topic. Both priors were symmetric. Please note doc_topic_prior will be interchangeably referred to as alpha and topic_word_prior will be interchangeably referred to as beta.

A higher alpha means documents are more likely to be a mixture of many topics, whereas a lower alpha means documents are likely to be made up of fewer topics. Similarly, a higher beta means the words that make up a single topic's vocabulary are likely to be present in a different topic's vocabulary. With a lower beta, the words that define a single topic's vocabulary are more likely to be unique to that topic.

In the first experiment, n_components was varied from 8-27 increasing by 1 per test. doc_topic_prior was varied from 0.02-0.2 increasing by 0.02 per test. topic_word_prior was varied from 0.02 to 0.4 increasing by 0.02 per test. Figures 2, 3, and 4 give an example of the testing process on the post-covid dataset.

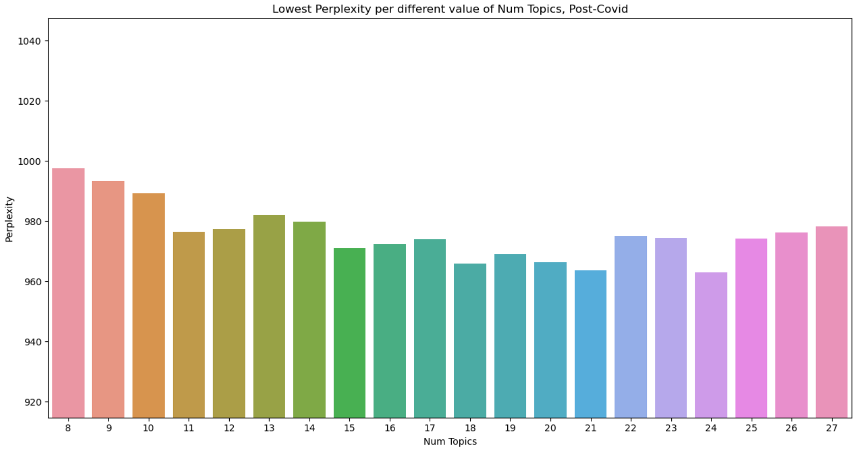

Figure 2: Lowest value of perplexity per value of n_components

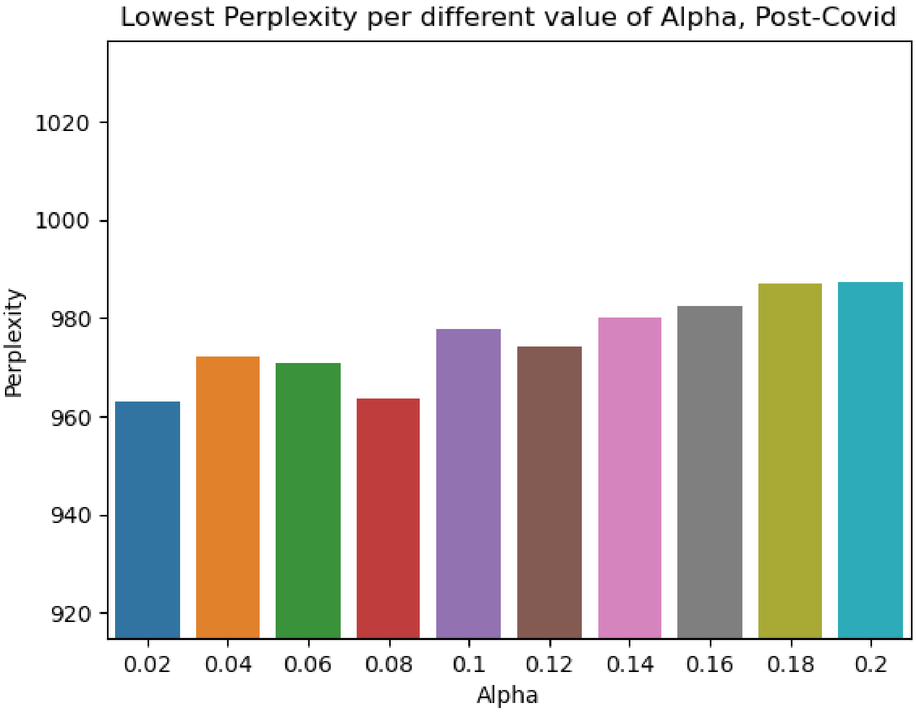

Figure 3: Lowest value of perplexity per value of doc_topic_prior

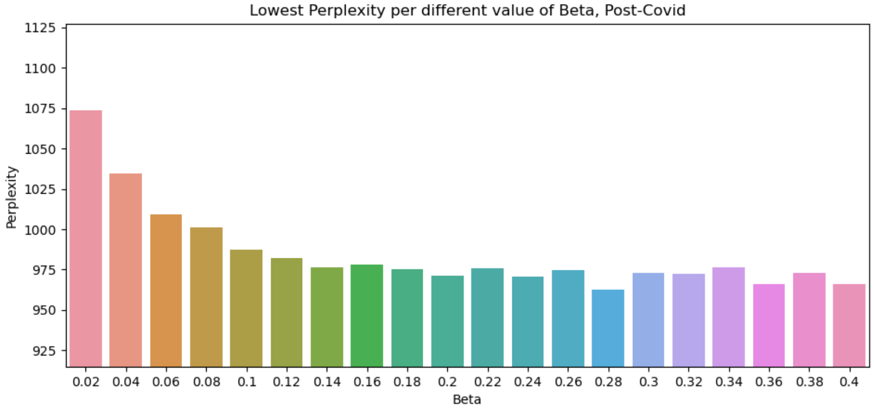

Figure 4: Lowest value of perplexity per value of topic_word_prior

The results show the best setting for parameter n_components was 18, and the best setting for parameter topic_word_prior was 0.28. After more testing, the best parameter for doc_topic_prior was found to be 0.01.

4 Results

Figure 5 shows the results of LDA after running on the pre-covid dataset, and figure 6 shows the results of LDA after running on the post-covid dataset.

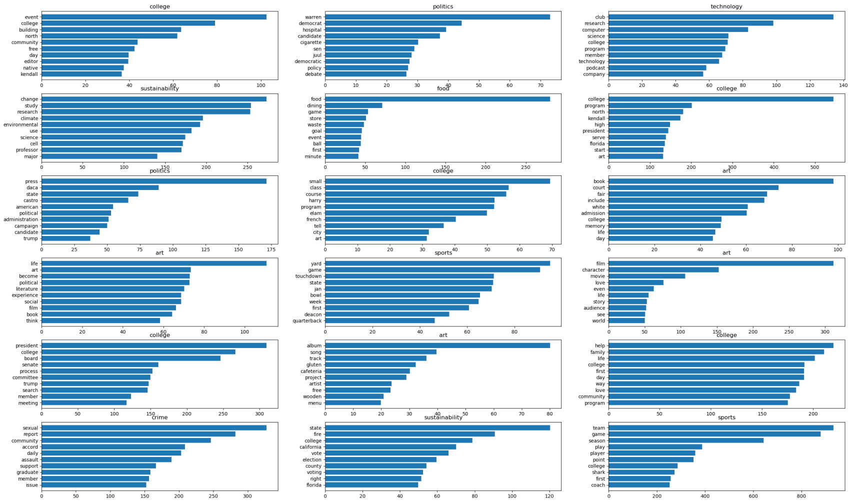

Figure 5: Words that describe individual topics, pre-covid dataset

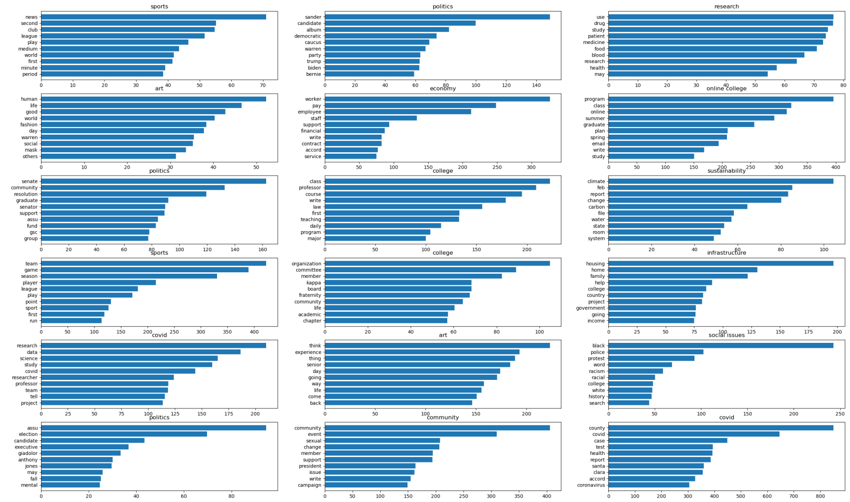

Figure 6: Words that describe individual topics, post-covid dataset

For pre-covid, the topics are politics, technology, food, art, college, sports, crime, and sustainability. For post-covid, the topics are covid, politics, community, online college, research, art, social issues, sports, college, economy, sustainability, and infrastructure. Topics such as college and art appeared multiple times, which is why there are a differing amount of topics for pre-covid and post-covid.

Figures 7 through 11 show the prevalence of post-covid topics in newspapers. Each bar represents the proportion of articles that pertained to the given topic by month.

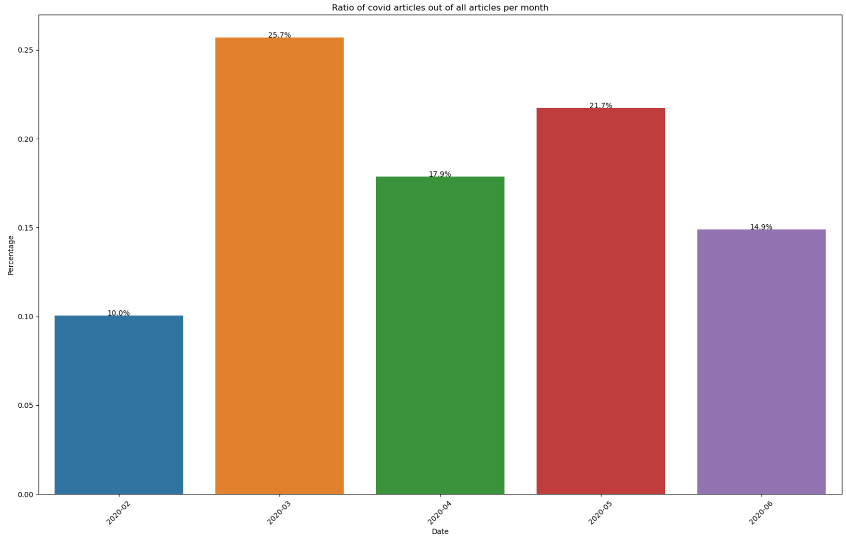

Figure 7: Ratio of all articles that pertain to covid per month

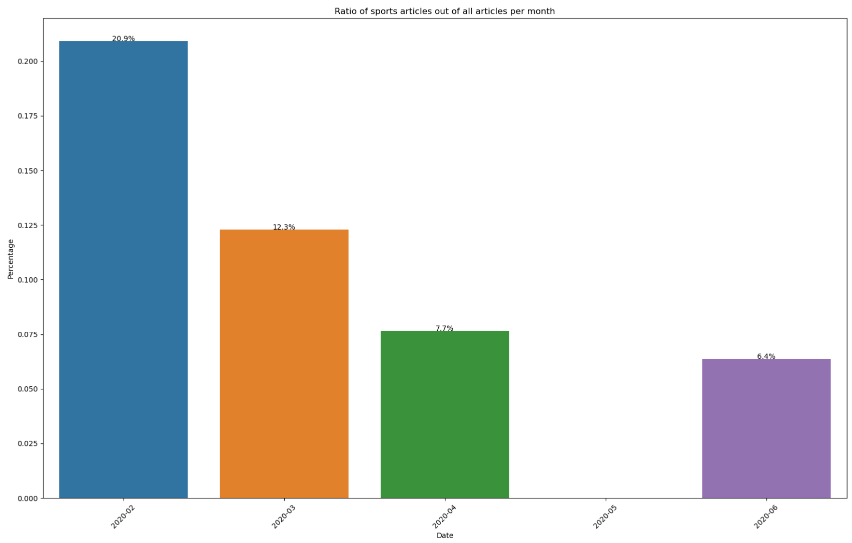

Figure 8: Ratio of all articles that pertain to sports per month

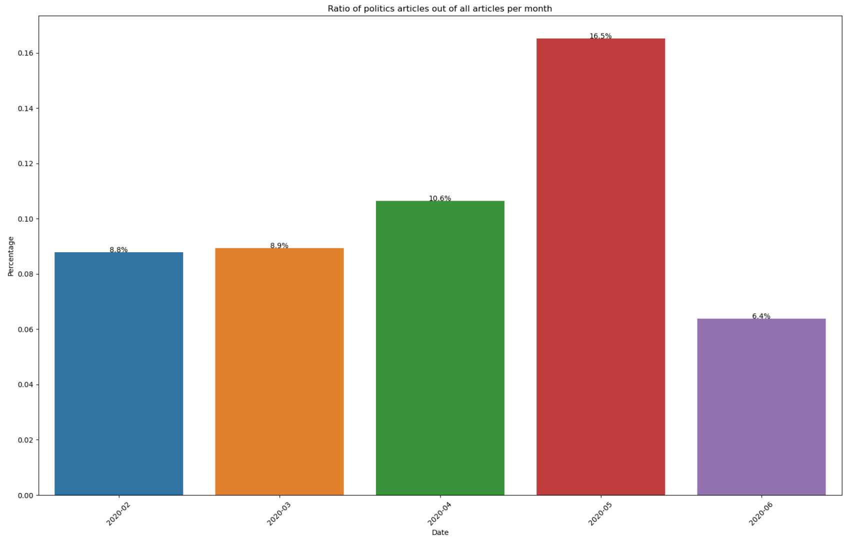

Figure 9: Ratio of all articles that pertain to politics per month

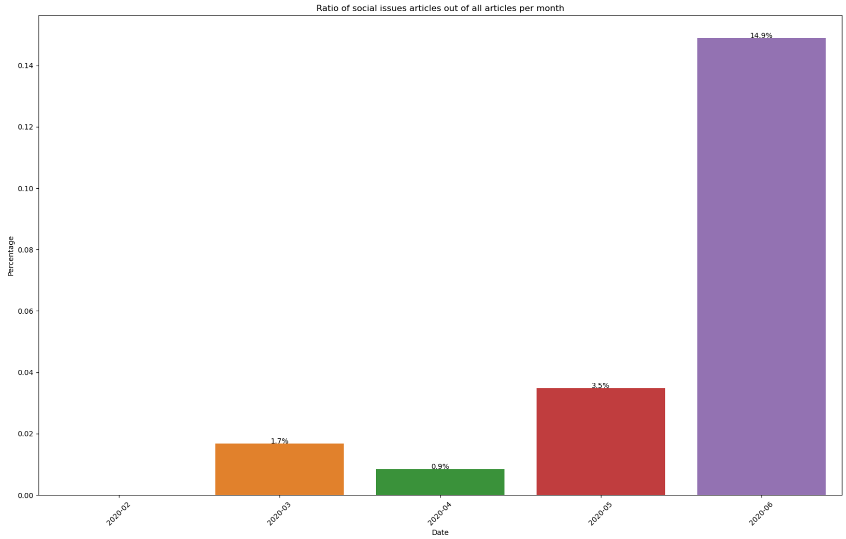

Figure 10: Ratio of all articles that pertain to social issues per month

Tables 2 and 3 show the results of sentiment analysis on each topic in both pre-covid and post-covid datasets.

Topic	compound	neg	neu	pos
politics	0.68943	0.086	0.745848	0.168
technology	0.735637	0.0701463	0.750659	0.179098
food	0.899372	0.07316	0.73276	0.1942
art	0.647632	0.103962	0.697624	0.198481
college	0.754758	0.0656401	0.747156	0.187224
sports	0.803516	0.0825299	0.715478	0.201925
crime	0.158632	0.111042	0.747617	0.141308
sustainability	0.666163	0.076	0.749063	0.174908

Table 2: Sentiment analysis of pre-covid topics

VADER sentiment analysis shows four different statistics: compound, negative (neg), neutral (neu), and positive (pos). The results for statistic compound are in range -1 to 1. Compound shows the overall polarity of a text; -1 indicates extremely negative sentiment, 0 indicates neutral sentiment,

Topic	compound	neg	neu	pos
covid	0.436869	0.0435486	0.881049	0.0753819
politics	0.585132	0.0500595	0.855667	0.0942976
community	0.572771	0.0503053	0.844705	0.104979
online college	0.668029	0.0322405	0.90343	0.0643544
research	0.461312	0.0623462	0.841308	0.0963462
art	0.814507	0.0545745	0.811766	0.133638
social issues	-0.517912	0.0864375	0.843125	0.0703125
sports	0.789883	0.0467312	0.84143	0.111849
college	0.775758	0.034481	0.866266	0.0992405
economy	0.564	0.0521667	0.870767	0.0771667
sustainability	0.0526231	0.0754615	0.851846	0.0727692
infrastructure	0.843922	0.0463265	0.848531	0.105102

Table 3: Sentiment analysis of post-covid topics

and 1 indicates extremely positive sentiment. The other 3 statistics are in range 0 to 1; they show how much of the text is made of negative, neutral, and positive words. For instance, 0.043 in the 'neg' column means approximately 4.3% of the text had a negative polarity.

5 Discussion

The results exemplified in figures 5 and 6 show that there were many overlapping topics between the pre-covid and post-covid datasets. However, the implications of covid were visible in the additional topics 'online college' and 'covid' produced from the post-covid dataset. Figure 7 shows that the coronavirus weighed heavily on writers' minds; even in February, when the virus had not prompted universities to send students home, 10% of the articles were related to the virus. Beginning in March and continuing to June, the covid topic was very prevalent in the news, ranging from 15% of produced articles per month to 25% of produced articles per month.

Another interesting development, the decline of sports, is shown in figure 8. As the pandemic started to take hold in the United States, the percentage of articles about sports decreased from 20% in February to 0% in May. By this point in time, students had been sent home, and no college athletics remained functioning. However, the percentage of sports articles went from

0% in May to 6.4% in June. The author believes this change is indicative of people becoming accustomed to life in the pandemic. During the initial shutdown in the United States, it was impossible to have any sort of organized sports due to risk factors such as heavy breathing, and having people in close proximity without masks. Recently though, the NBA found a way to isolate players and host games in 'the bubble'. Thus, the author believes the reemergence of sports into college news shows how adaptable humans can be even when faced with great setbacks like the covid pandemic.

The most intriguing visualization to examine is figure 10, the ratio of all articles that pertain to social issues per month. The keywords that define 'social issues' as a topic are: black, police, protest, word, racism, and racial. There was an explosion of articles about this topic in June - in March, April, and May, the topic appeared as 1.7%, 0.9%, and 3.5%, respectively. However, in June, this topic accounted for 14.9% of all articles written. The author believes that the primary catalyst for this change was the videotaped murder of George Floyd. However, Floyd's death was not an isolated incident. In 2016, Alton Sterling was murdered; in that same year, Philando Castille was murdered; and in 2017, Eric Gardner was murdered, all at the hands of police. Each of these incidents were recorded on body cam footage, and for viewers, it was clear to see that each of these killings were unjust.

There was a notable difference between each of these events. After George Floyd, there was unprecedented change in how people reacted; for Alton Sterling, Philando Castille, and Eric Gardner, people seemed to take notice for a little while, but there was no sustained societal action taken against police. Presently, societal reaction appears different - people are still posting about George Floyd on social media, many prominent celebrities and noteworthy people voiced their anger, and numerous riots broke out throughout the United States. The uptick in articles about social issues demonstrates this well. The author's theory is that covid has interrupted the daily lives and normal distractions that allowed people to forget about these murders. All the deaths were clearly sickening and unjust, but issues at work, school, or family life would arise, and people had to prioritize their own problems rather than pushing for societal change. Since covid has forced everyone to stay at home and isolate, people are not as easily distracted anymore; everyone has had more time to ponder how these killings could stop, and what they could do to help. Additionally, while people were stuck at home, many felt that their freedom to move was restricted, potentially leading to an increase in empathy for minority groups whose rights have been ignored.

Thus, the author's hypothesis for this change is: due to isolation, a lack of normal distractions, and the restricting of civilians' rights to move freely, people were drawn to be more active in combating social issues that regular life otherwise distracted them from.

Another intriguing aspect of the 'social issues' topic is the VADER sentiment score. Although most other topics were scored as positive overall, the social issues topic had a compound score of -0.51, indicating the polarity of these articles were negative (table 3). Thus, the way authors wrote the newspapers demonstrates the civil unrest and anger that many were feeling after George Floyd's death. Not even articles about crime in the pre-covid section were scored as negative, rating at 0.15 (table 2). Interestingly, the sustainability topic had the second most negative polarity out of all topics; it was not negative, but at 0.05, indicating overall neutrality.

In order to improve the discoveries made in the paper, more articles produced in June, July, and August could be scraped and run through testing. This would give further insight as to how people adapt to the limitations that the pandemic has forced upon them. Although many have gotten over the initial shock and change that the pandemic started, more recent articles could shed light on how people adapted to the new world around them. An additional way to improve discoveries is by using a metric other than perplexity, such as topic coherence. Using both perplexity and another metric would help validate the accuracy of the best possible parameters.

6 Prior Work

The main prior work that this paper relies on is topic modeling. Topic modeling algorithms are designed to input a large corpus of documents and output a list of topics that can be detected throughout the corpus. The main algorithm Latent Dirichlet Allocation (LDA)[BNJ03] is designed to group words that are often seen together. Topic modeling algorithms are unsupervised; thus, there is no 'ground truth' in which the programmer can use to validate the results of the algorithms. Other common topic modeling algorithms are Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA)[Dum+88]. LDA, LSA, and NMF all rely on a Bag-of-Words representation of documents. Each document is considered to be an accumulation of the words inside, and the order of when words appear is ignored.

There are many variations of the LDA algorithm, including supervised

LDA (sLDA) [BM07], discriminative LDA (discLDA) [LSJ09], and maximum entropy discrimination LDA (medLDA) [ZAX12]. However, these algorithms are not readily available through common Python modules such as Gensim or Sklearn. While papers exist online that explain how these algorithms work, they are not easily applicable to a project, which is why LDA was the main contribution towards the discovery of topics in newspapers.

7 Conclusion

College newspapers shed light on the feelings of the general populace. The change in topics from pre-covid to post-covid demonstrates how people initially reacted to the pandemic, and how the consequences of the pandemic led to an increased need for societal change. It shows how daily life changed, including the decline in articles about sports. This project demonstrates how topic modeling algorithms used on college newspapers can shed light on the societal changes that the COVId-19 pandemic brought about.

8 Acknowledgments

I would like to first thank Dr. Natalia Khuri for her continual support, guidance, and patience with the members of the NLP group. Her expertise and advice assisted me greatly throughout the summer, and I could not have created this without her help. Thank you to the other members of the NLP group, Sarah Parsons, Tian Yun, and Josh Mannion for their support - even though we were all far away, I felt a strong sense of community in meetings with them, and their contributions helped me get here. I would like to extend my gratitude to Cody Stevens for the assistance with the execution of computational experiments. I also acknowledge the Distributed Environment for Academic Computing (DEAC) at Wake Forest University for providing HPC resources that have contributed to the research results reported within this paper. URL: https://is.wfu.edu/deac

References

[Dum+88] S. T. Dumais et al. “Using Latent Semantic Analysis to Improve Access to Textual Information”. In: *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems. CHI '88. Washington, D.C., USA: Association for Computing Machinery, 1988, pp. 281–285. ISBN: 0201142376. DOI: `10.1145/57167.57214`. URL: `https://doi.org/10.1145/57167.57214`.

[LB02] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit". In: *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*. 2002.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[BM07] David M. Blei and Jon D. McAuliffe. "Supervised Topic Models". In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. NIPS'07. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007, pp. 121–128. ISBN: 9781605603520.

[Ric07] Leonard Richardson. "Beautiful soup documentation". In: *April* (2007).

[LSJ09] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. "DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 897–904. URL: `http://papers.nips.cc/paper/3599-disclda-discriminative-learning-for-dimensionality-reduction-and-classification.pdf`.

[Ped+11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[ZAX12] Jun Zhu, Amr Ahmed, and Eric P. Xing. "MedLDA: Maximum Margin Supervised Topic Models". In: *J. Mach. Learn. Res.* 13.1 (Aug. 2012), pp. 2237–2278. ISSN: 1532-4435.

[Yan+19] Shuo Yang et al. “Unsupervised Fake News Detection on Social Media: A Generative Approach”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 5644–5651. DOI: 10.1609/aaai.v33i01.33015644.