UNIVERSITÄT
DES
SAARLANDES

Department of Language Science and Technology

Muhammad Umer Butt

Computational Linguistics

<u>Analyzing Trends in US News</u>

Project Report

26th March - 2023

A note before we start:

As per the feedback, I modified my initial idea and worked on a new project that I believe is both interesting and relevant to the topics we studied in the lecture.
To determine the feasibility of my new idea, I consulted with other students from this and previous years to gain insight into what kinds of projects have been successful in the past. After much consideration, and trying out different projects. For instance, I attempted to work on n-gram language generation for an Urdu news dataset but the project faced several obstacles and resulted in bad-quality output due to language structure and lack of techniques to process it. Also, I wanted to stick with LDA as it was the most intriguing topic discussed in class, and I am passionate about developing my skills in this area. Furthermore, I intend to pursue a career in data analysis, and expertise in LDA (and topic analysis will undoubtedly be a valuable asset). So, I decided to go with the analysis of trends in the US news dataset of HuffingtonPost[1]. The whole project code, model files and output images are stored in the github repository [2]

Introduction:

News papers have long been a primary source of information for individuals seeking information about the current events happening around the world. But over time, the medium of news has changed a lot. After the start of the era of digitalization, news websites for the most part, have taken the place of newspapers. This is because of the fact that news websites are much more timely than their print counterpart. People do not want to wait till tomorrow for today’s information and that is exactly what news websites provide. Within minutes of the event, it's up on the news website where hundreds of people can read about it.

Since it's easy to publish news on websites now, there are a plethora of news articles online. Websites have these articles divided into a variety of categories For example local news, world news, sports etc. This in turn, has led to many possibilities of analyzing the news articles for identification of the trends and public opinion over the years. This can be quite helpful in understanding the political and cultural situation in the country or even the world overall.

To this end, this project utilizes a newly available dataset of Huffington Post articles, more commonly known as the “HuffPost category dataset”[1]. This dataset comprises approximately 200,000 articles' headlines and short descriptions, spanning from year 2012 to 2022. By utilizing a range of linguistic techniques such as Latent Dirichlet Allocation (LDA) and Named Entity Recognition (NER), the project seeks to identify latent topics and trends in the HuffPost dataset. By doing so, I aim to offer new insights into the shifting trends and patterns in news articles over the years.

Project Justification

Analyzing the newspaper is useful in determining the trends and patterns of events that have happened over time, both in terms of news topics themselves and the way they are reported.By using the tool like Latent Dirichlet Allocation (LDA) and Named Entity Recognition(NER), we can identify the latent topics that may not be apparent without such analysis. These insights can help understand the public's perception and identify the areas of concern or interest.

In the context of this project, I have used a recently released dataset for analysis. Although the author of the dataset did provide some very basic analysis but this project provides an in depth analysis. Thus helping to provide a comprehensive understanding of the data which can serve as foundation for for future work.

Related Work:

The availability of data has been a huge factor in enabling the research work in the field of data science and machine learning. And there are indeed some datasets of news articles but the availability of a dataset which is big enough to span over the years, can be updated, and is freely available is not a small ask. The recently released Huffington Post dataset, commonly referred to as the "HuffPost category dataset" is one such example of a potentially valuable resource for researchers.

Since this is a relatively new dataset, released in October 2022, there hasn't been a lot of work on this dataset yet. The author of the dataset did provide a very simple analysis[1], For example, the count of articles in each year or the headline's length across the years but that is it. Compared to that, in this work, I do a detailed analysis using LDA, NER and other techniques.

There have been other datasets of news as well to my knowledge, But this is the most recent and one of the biggest datasets, in terms of the number of articles. Also, this dataset can be further Updated as HuffingtonPost still keeps the data on their archive[3] and hence it can easily be crawled. That archived data is allowed according to site rules mentioned on robots.txt for the website (robots.txt mentions what is allowed to crawl and what is not). Furthermore, I have also contacted the author of this dataset in regards to expanding this dataset to make it more generic so anyone can just run the script and get the latest data. Considering all this, this is a very good dataset and other research might find it useful for other tasks too for example sarcasm detection in the news or fake news.

Data:

The Huffington Post dataset was curated by Misra[1] at the end of 2022. It consists of around 200,000 news articles' headlines with supported short descriptions of the article. The articles are from January 2012 to September 2022. Even though the dataset contains articles from all of these years, the number of articles varies for each year. Each article is assigned a category, for example, Politics, Comedy or Wellness. There are 42 categories in total in the dataset but as we will see later, some of the categories can be combined together because they represent the same entity. One thing to note is that there are more articles for 2012 to 2018 than for 2018 to 2022. To be precise 2012 to 2018 have 200,000 articles whereas 2018 to 2022 have only 10,000 articles. This is also shown in fig 1. As mentioned earlier, each article is assigned a unique category. And as expected some of these categories appear more than others. I also had to combine some of the categories as the labels were not correct, for example, world news was categorized with "WorldPost" and "The WorldPost". So, I combined them into one category. In figure 2, you can see the total count of each category throughout the dataset after the preprocessing step. Later I analyze this in detail by going into each year and see how the number of articles changes over the years for a specific category.

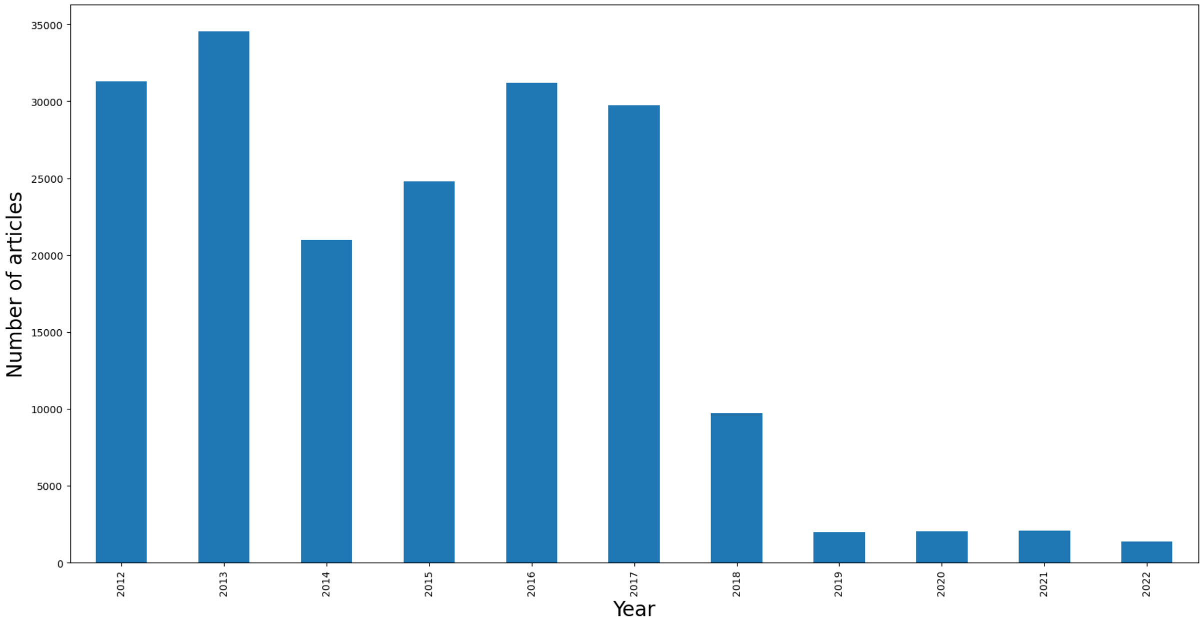


Figure 1. Number of articles over the years in HuffPost dataset

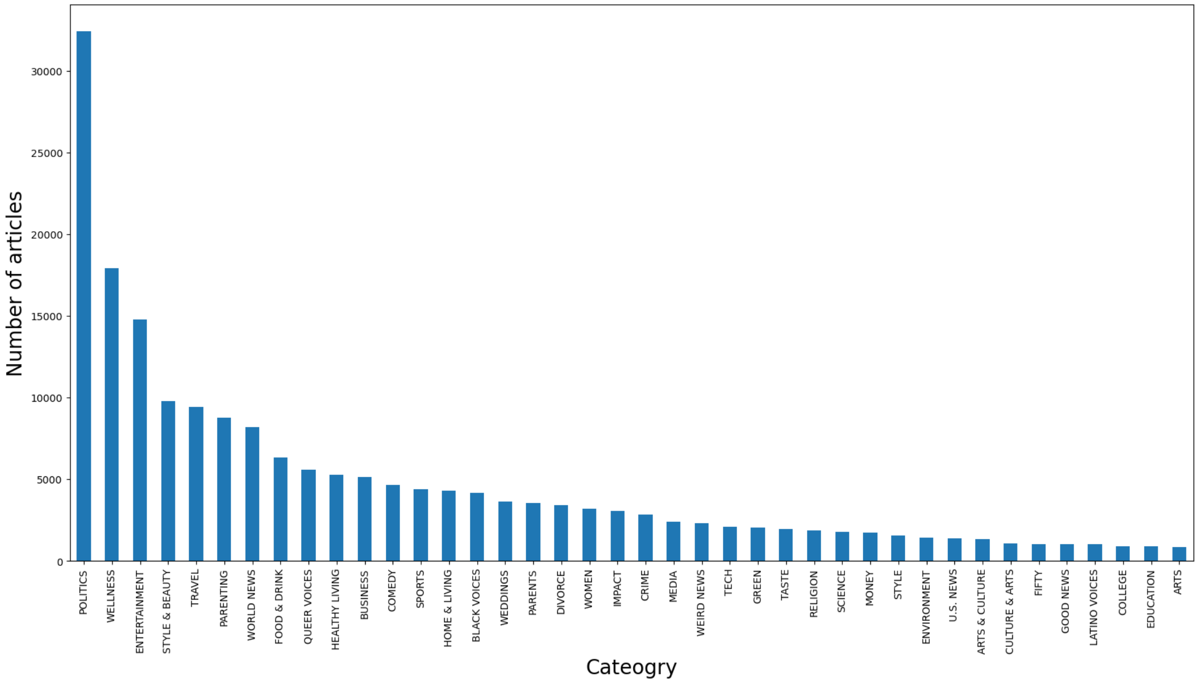


Figure 2. Number of articles by category

Analysis of Data:

To analyze the data in detail, first I concatenated the headline and the short description of the article because I wanted each article to have as much detail as possible. I also cleaned the data by removing the punctuations and stop words. I used two techniques LDA and NER for analysis. Although the articles were already categorized, I wanted to use LDA for further analysis within each category. Especially politics since not only politics has the most number of articles in the dataset but also because political news seems more important. I used NER to extract the personalities most discussed in the dataset and especially in political articles for the same reasons mentioned above.

Variability of categories over the years

Before we dive into LDA and NER results, I am going to discuss some general analytics. For example how the number of articles within a category vary over the span of the 10 years. Each article is associated with one category and there are around 40 categories in total (after pre-processing). In figure 3 and 4, you can see the top most categories from 3 years; first year of the dataset(2012), mid year (2017) and last year of the dataset(2022). As previously mentioned, you can see that since the number of articles will decrease after 2018. So, naturally the number of articles in the 2022 graph is less than 2017 or 2012, and that is also the reason why I have shown these three years in separate graphs because otherwise due to the scale of number of articles in 2017, the graph of 2022 was getting too small. Another thing to note here is that there are no articles categorized into politics for the years 2012 and 2013, that is a limitation of the archival process.

What these graphs tell us overall is what interests news readers over the years because news agencies publish more of what they think the individuals will be more interested in. So, for example as we can see, politics has gained a lot of attention through the years but at the same time, less material has been published on general well being i.e wellness articles. It's also noticeable that even though the number of articles in 2017 are much more than in 2012, still the wellness articles decreased from 2012 to 2017. The articles on “Black voices” also increased from 2012 to 2017, and compared to 2017, their rank in the year 2022 is lower. This is probably because of the fact that in 2017 this topic was on the rise in the US. Entertainment as we can see has remained relevant through the years and that is probably because of the huge boom of Hollywood and also a general interest by the American public. The sports articles are also discussed more than the other topics as the years go and that shows that people are more keen towards reading about this topic. The category of Travel decreased from 2017 to 2022 and that is probably of the pandemic and less ppl being able to travel after 2020. But we may see a rise in it again once we get the full data of 2022.

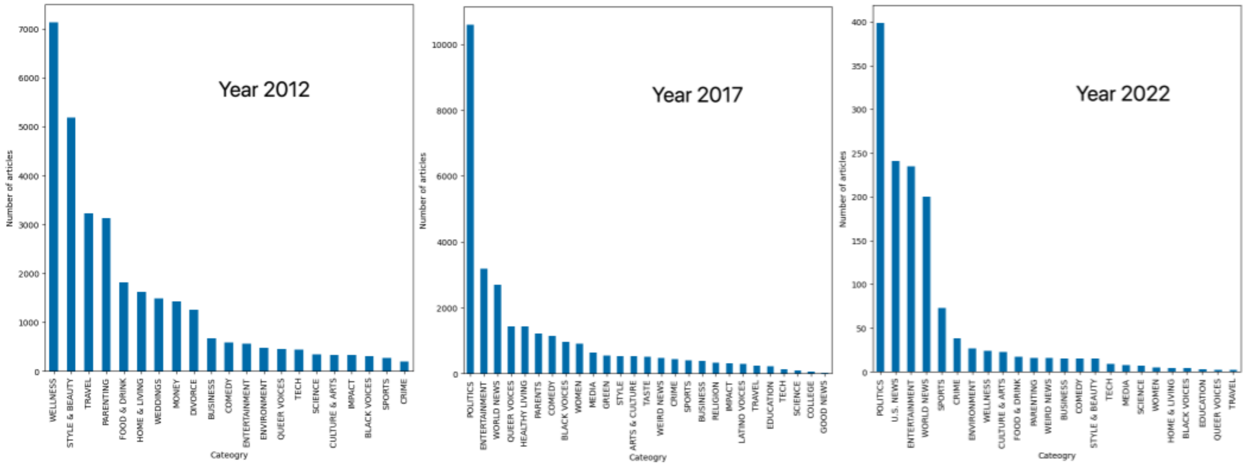


Figure 3. Category wise distribution of number of articles over the years

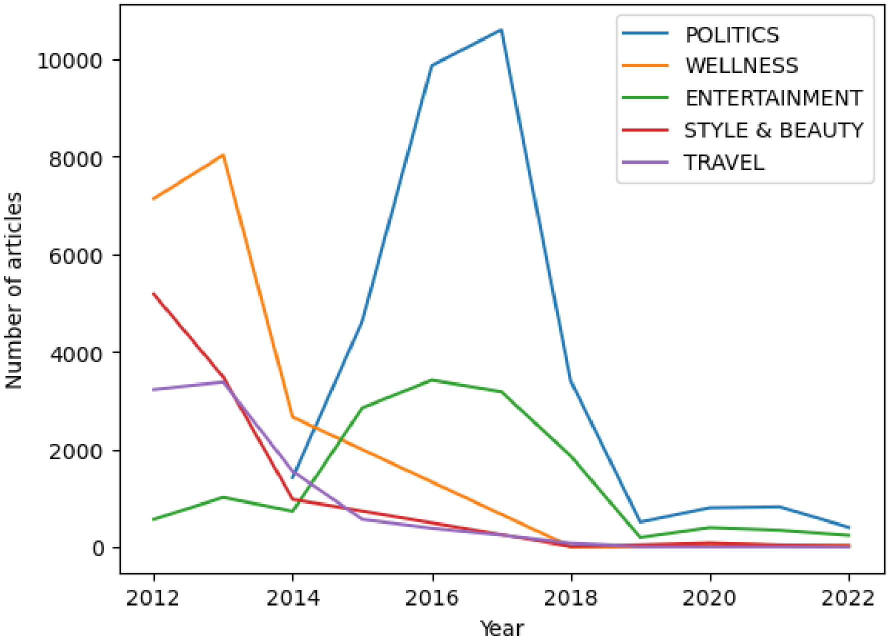


Figure 4. Counts of top 5 categories over the years

Personalities discussed over the years

In this section, we discuss the personalities discussed most in out dataset.
For this, I tried Spacy and NLTK's NER packages. Spacy has 4 different models, small, medium, large and transformers based. I got the best results using a spacy transformers based model. For NER I kept the data in the actual case-form because the NER models work better if the named entities and especially names' first letter is capitalized. Once the entities were detected then I had to map all the different names of a single person to one name. It was mostly due to typing errors but this processing made the results more robust.
I did this for the top 7 most occurring personalities.

Coming to the results, as you can see in fig 5, Donald Trump has been mentioned way more than any other personality across the years. In fig 6, I show two charts because Donald Trump was skewing the data too much. We can clearly see from these graphs, how Trump's popularity was at the peak in 2015 to 2018, that is before he was elected as the president in 2017. started declining around 2018 and 2019, but around the end of his term, the news articles started talking about him again and I believe that it is because the political situation around 2019-2021 was very dynamic in the US and elections were near too. We also see a clear rise of joe biden mentions starting from 2019 elections and as we know he became the president of US at the start of 2021. One thing to note here is how around 2017 to 2018 there is a constant drop across all personalities. This is probably because of the dataset as we have less number of articles after 2018.

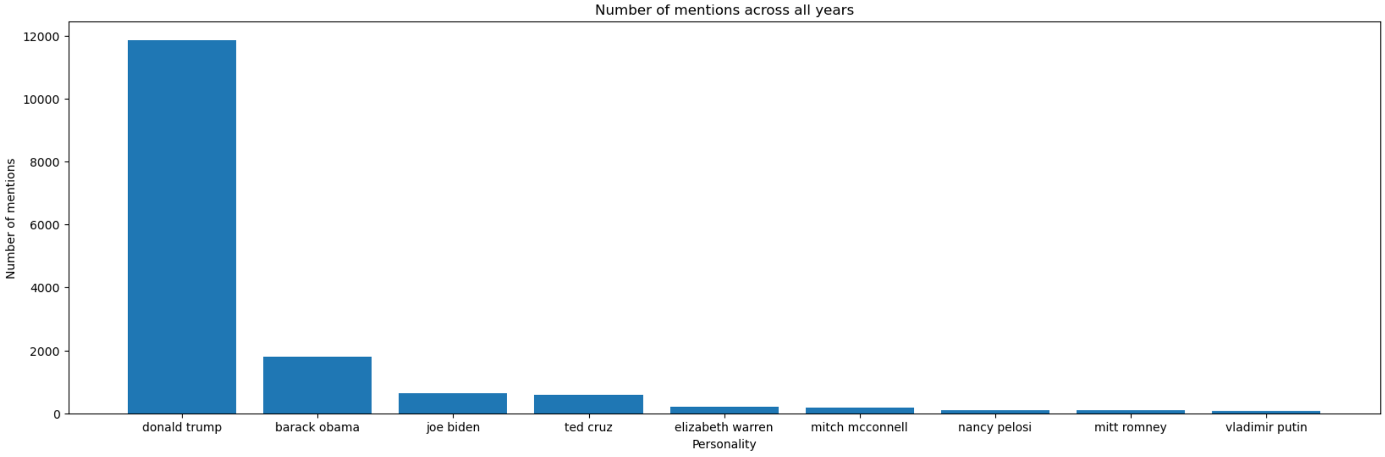


Figure 5. Total Number of mentions of personalities

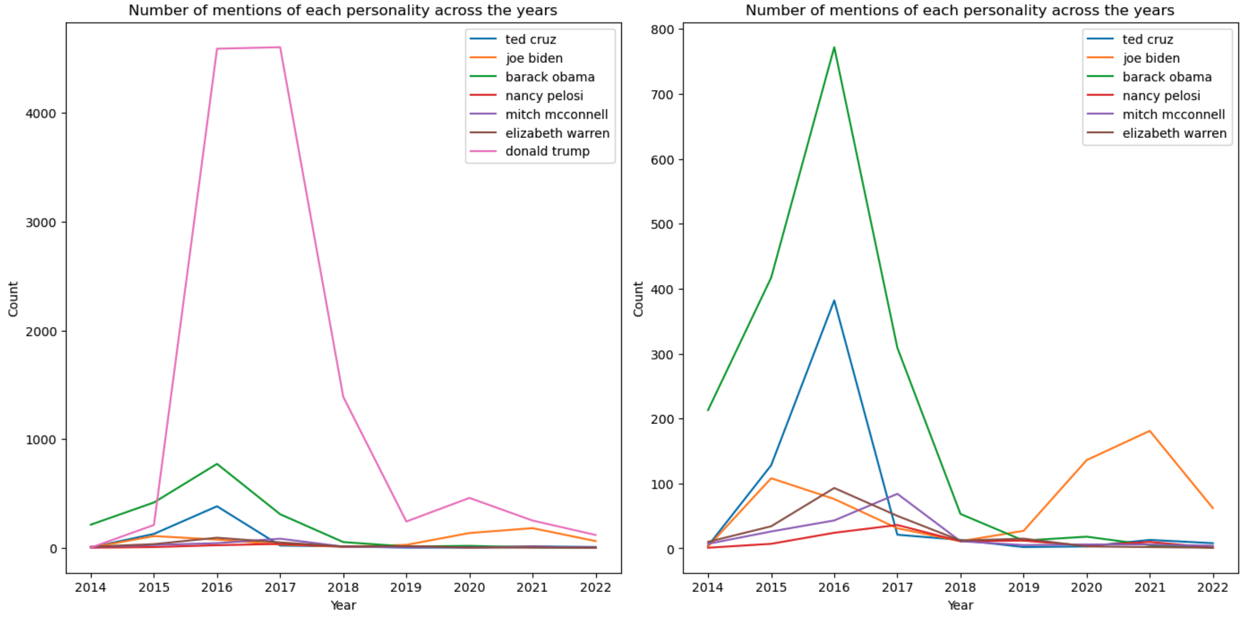


Figure 6. Number of mentions of personalities across years

Topic modeling:

In this section we discuss the finding of LDA analysis. To explore the topic modeling results, I applied LDA in three different ways. First, I applied it over the whole dataset spanning a decade. Second, I conducted the year by year analysis of the data. Lastly, I focused on two most occurring topics, Politics and wellness and applied LDA on their data for each year separately. In summary, I obtained four different LDA results. LDA on the whole dataset, LDA for each year, LDA for each year of politics and LDA for each year of wellness.

For LDA I used my own implementation of LDA which I made for LDA assignment. I found the same parameters of 0.02 and 0.1 to have generated satisfactory topics. I tried other alpha and beta +-0.2 values for both but 0.02 and 0.1 generated better results. I ran 500 iterations for the whole dataset. Whereas for each 'year-by-year' and 'category-year' experiments I ran LDA for 100 iterations. Similarly I ran LDA for 20 topics when running for the whole data while 5 topics when running for year-by-year and category-year experiments. These choices were made due to the smaller amount of data when working on a subset.

Moving on to the results, First we discuss the topics extracted from the LDA run on whole dataset together. As mentioned above, I ran it for 400 iterations and for 20 topics. Some of the topics were significant. For example the topic 6 as shown in fig7, represents the topic of equality and justice, which has gained quite a lot of coverage in the last decade. Similarly in topic 16 figure 8, we can say that the topic is about police action. Which has been major topic of discussion in the US due to police brutality, shootings and crimes.

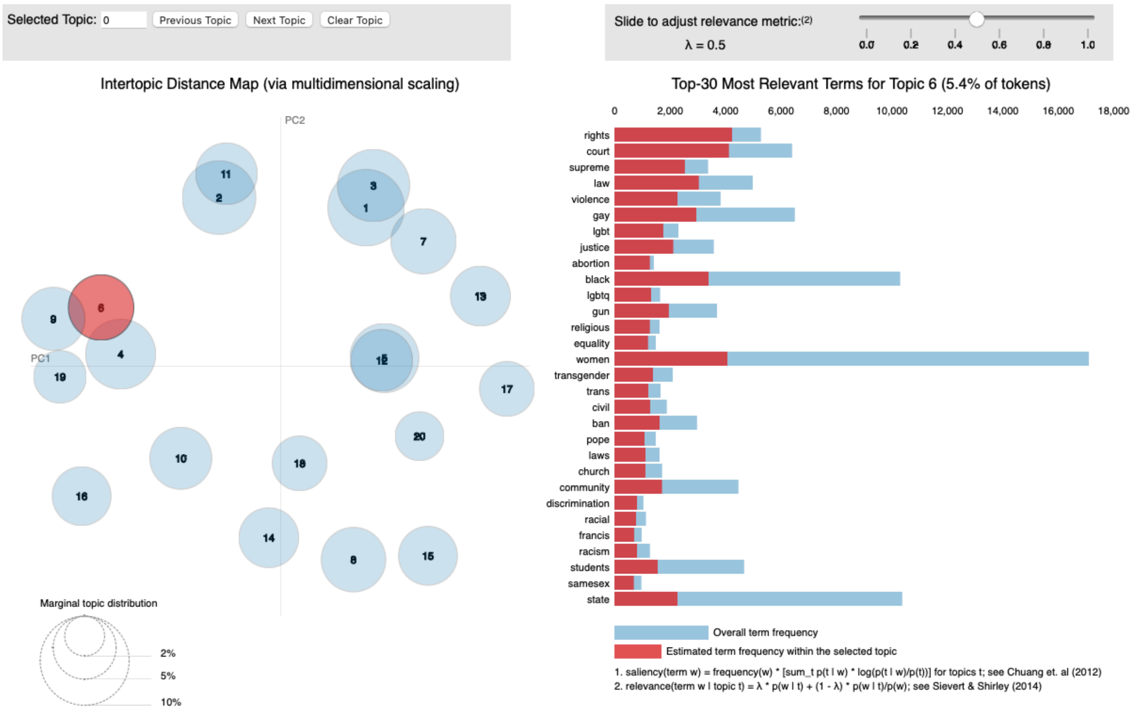
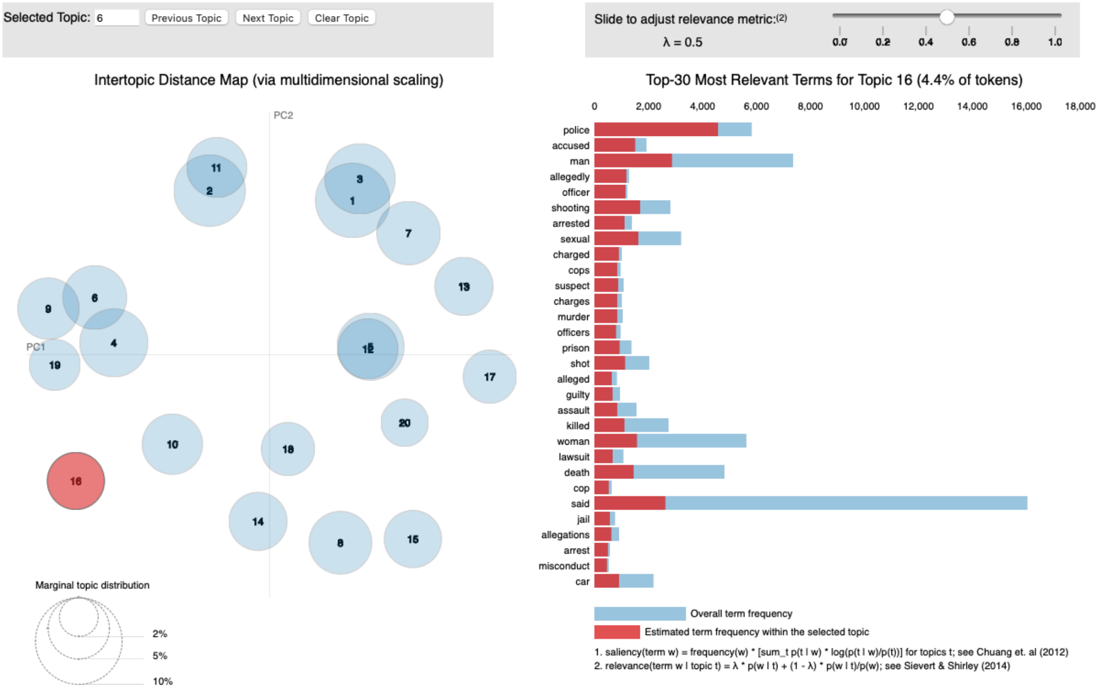

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.5
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
11
2
3
1
7
13
9
6
4
19
12
15
17
20
10
18
16
14
8
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 6 (5.4% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
rights
court
supreme
law
violence
gay
lgbt
justice
abortion
black
lgbtq
gun
religious
equality
women
transgender
trans
civil
ban
pope
laws
church
community
discrimination
racial
francis
racism
students
samesex
state
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 7. A topic (6) from LDA run on the whole dataset. Topic probable name :equality and justice

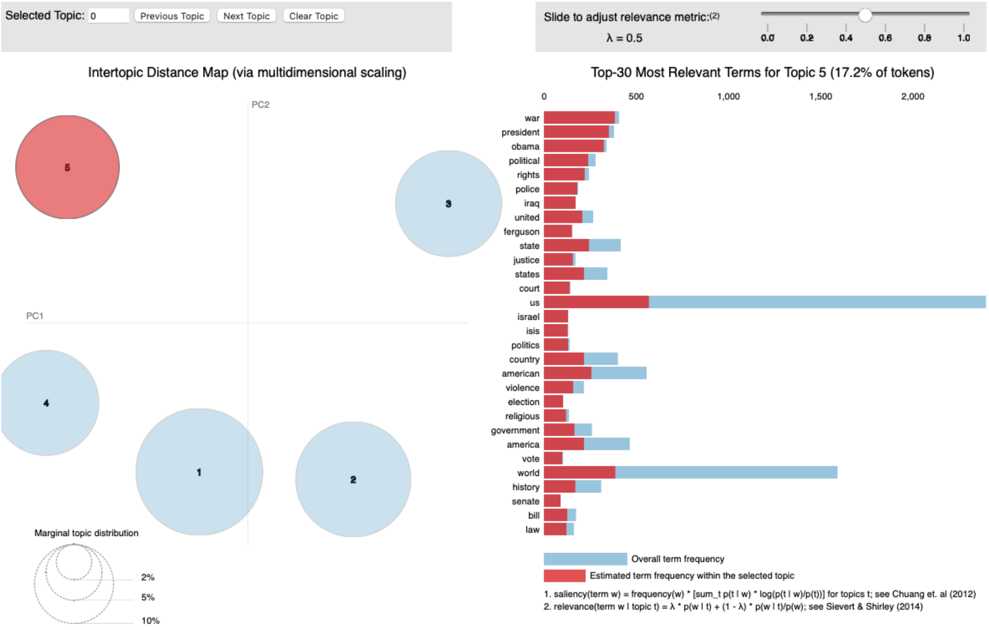


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. A topic (6) from LDA run on the whole dataset. Topic probable name : crime reporting

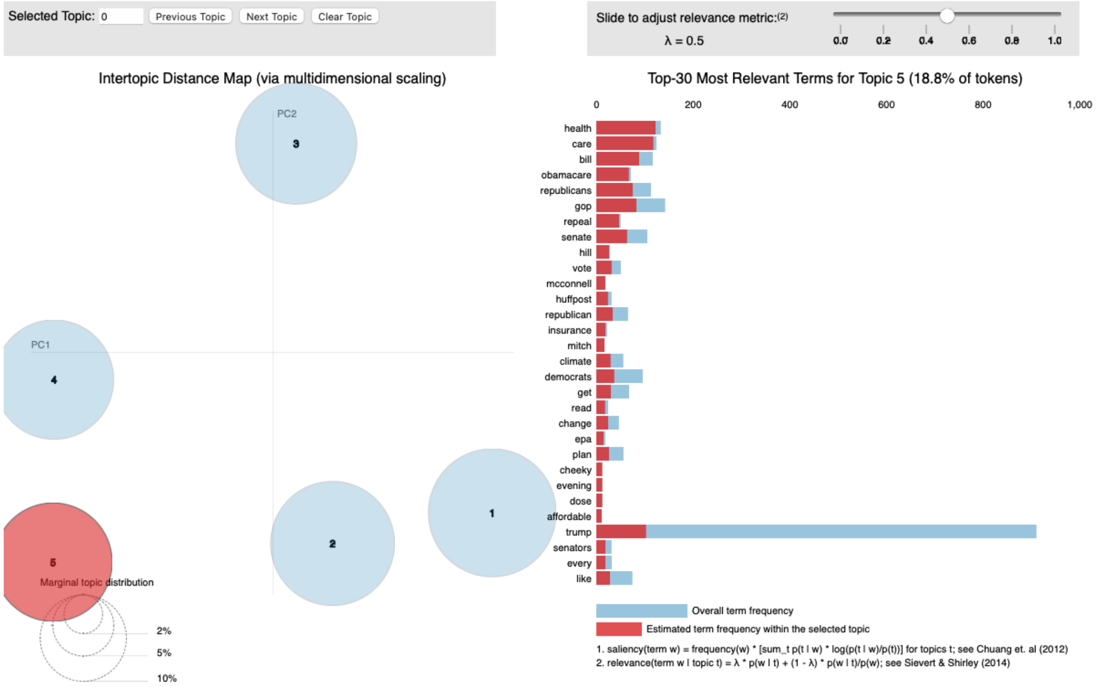
Next, let's move to the year by year analysis using the LDA. As mentioned earlier, I trained LDA for 100 iterations and for 5 topics. Running for 500 iterations for each of these sub experiments would have been too time consuming. Nevertheless we got good results with 100 iterations.. For example topic 5 in fig 9, shows a topic related to war, Iraq, obama, isis and election. Probably because the elections were approaching, the US was debating ending the wars and Obama was the president and his policies for war were questioned at that time. On the other hand if we look at the 2017 data, figure 10, we realize this topic is about obama health care and 2017 was election year and health care packages were indeed discussed by democrats but at the same time, republicans were on the rise and as we know republicans(Trump) won in 2017. If we compare this with the 2022 data (fig 11), we see that the focus has completely shifted towards the Russia-Ukraine war. Biden is in this topic probably because of his policy discussions, gas is mentioned because it has become a problem due to this war, Putin and Zelenskyy are in this topic because they are two heads of states involved in this conflict..

Overall, we see that the trends of topics shifted from US middle eastern war and its policies to health care packages and then towards republican policies and coronavirus as a topic in 2020 (i haven't attached the visualization of 2020 hare, but i will upload it on my repository) to finally in 2022 topics like russian invasion of Ukraine. This LDA analysis was helpful in understanding how the focus of news websites shifted from one to another over the years.



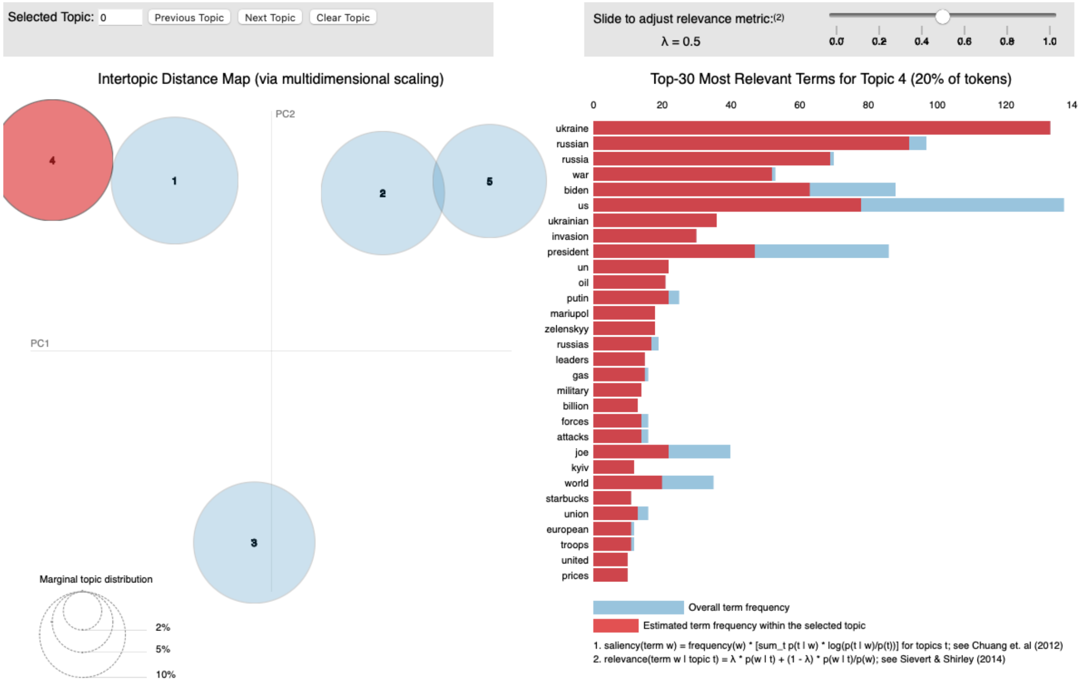
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9. Topic 6 of LDA on 2014 data. Topic probable name : US policies for War



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10. Topic 5 of LDA on 2017 data. Topic probable name : Obama health care package



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 11. Topic 4 of LDA on 2022 data. Topic probable name : Russia Ukraine war

Future Work:

This work lays a foundation for future work in the direction of analysis of news data. Since the data source itself is also expandable, by scrapping the HuffPost archived data [3], the methods mentioned in this work can also be applied on ‘newer’ data. One can also use the LDA trained objects/models and further train them on either more data or for more iterations since this functionality is supported by the LDA implementation. For this purpose I have uploaded the LDA trained models on the github repository which is publicly available.

About Code and project repository:

I used an ipython notebook for this project because of practical reasons. I had a lot of model files and images in this project all of which I could not attach here, so, I have uploaded the whole project to my github repository [2]. Although the code for every thing I have mentioned in this report is in the notebook, I have removed the repetitive lines, for example plotting the graph for multiple years and I have kept only one line which has to be tweaked only a little bit to use it for your desired year or any kind of category. I have also divided the notebook into sections to make it easier to read and understand.
project directory structure is mentioned in the readme file

References:

1) Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv:2209.11429 (2022).
2) https://github.com/UmerTariq1/Analayzing-topic-trends-in-US-news
3) https://www.huffingtonpost.com/archive/