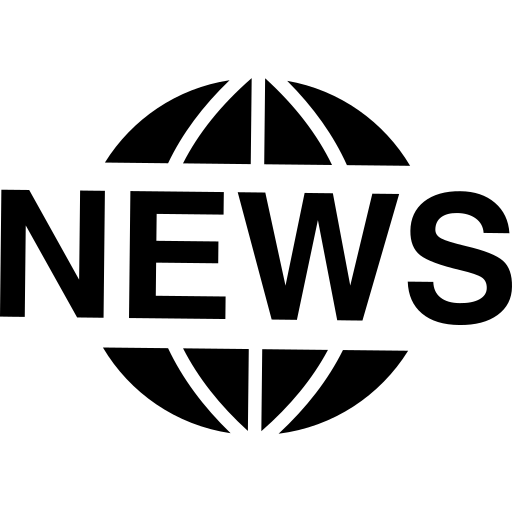
**A Report on**

**NEWS ARTICLES SENTIMENT ANALYSIS**

**AND TOPIC MODELING**

****

Submitted By

**Davis Silverman, Jianghui Li, Shian Li, Yashasvi Singh**

Submitted To

**Professor. Lu Xiao**

**May 8th, 22**

**Abstract**

Sentiment analysis is one of the widely used natural language processing techniques, particularly in the classification approach for extracting the content with emotions and attitudes. In such a manner, sentiment analysis may thus be thought of as a method for quantifying qualitative data using score markers. Despite the fact that sentiments are mostly subjective, quantitative analysis of emotions has many useful practices. Because of recent breakthroughs in machine learning and deep learning, sentiment analysis algorithms have grown more efficient. It takes a long time to manually classify emotional words. In today's world, where digital data is constantly being collected. Looking for insights from the acquired data might become difficult and tedious. Topic modeling is mainly used for semantic analysis and text mining in Natural Language Processing. Topic Modeling was developed as a method for organizing, searching, and comprehending large amounts of textual mental material that derives hidden patterns in the corpus resulting in improved decision-making. The topic is a probability distribution in Topic Modeling, using all characters in the text as the supporting set, showing the frequency of characters in the topic; characters with high importance to the topic have a higher likelihood of occurring. Our group has used both sentiment analysis and topic modeling (LDA) techniques to analyze news articles so that we can know the overall emotions of the news and frequently reported topics.

*Keywords: news, sentiment analysis, topic modeling, LDA*

The Table of Contents for this report is listed below.

* Introduction
* Dataset Gathering
* Dataset Description
* Exploratory Data Analysis
* Sentiment Analysis

- *Interpretation*

* Topic Modeling

- *Interpretation*

* Conclusion
* References

**Introduction**

Throughout history, news plays an important role in our daily life. People obtain information through various media and news sites and get to know topics they are interested in or notable information that they need to pay attention to. New media companies focus on delivering news to either the general public or the target public. In our project, we get data from three new sites which are Breitbart News, Fox News, and Newsmax. All three news media companies deliver their news to the same target public. Fox News is one of the biggest news media companies in the United States. Breitbart and Newsmax are smaller companies compared to Fox News. We are interested in how these news sites shift main topics over time and the similarities or differences in topic selection during the same period of time for these three news media companies since they target the same audiences.

The goals of our project are mainly divided into five parts, which are data gathering, data-preprocessing, exploratory data analysis, sentiment analysis, and topic modeling.

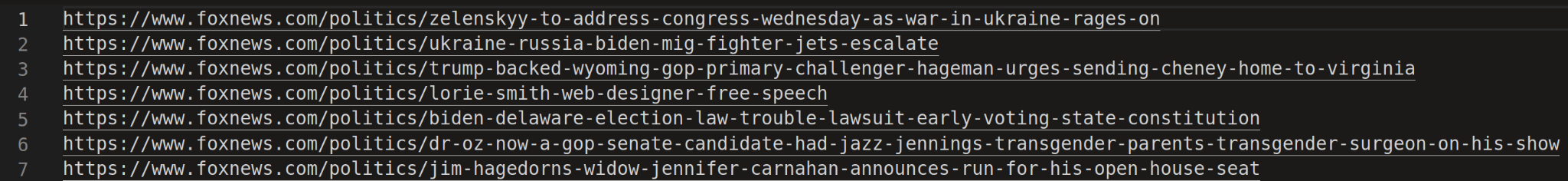
**Dataset Gathering**

We gathered data from three sources. The first source, Fox News, has data from December 29th, 2021 to March 14th, 2022. The data includes a wide range of topics in the political-sphere covered by Fox News. Breitbart, another popular conservative news site, has data from January 29th, 2022 to April 18th, 2022, and Newsmax from July 1st, 2021 to April 17th, 2022. These three sites cover a wide spectrum of conservative news-watchers.

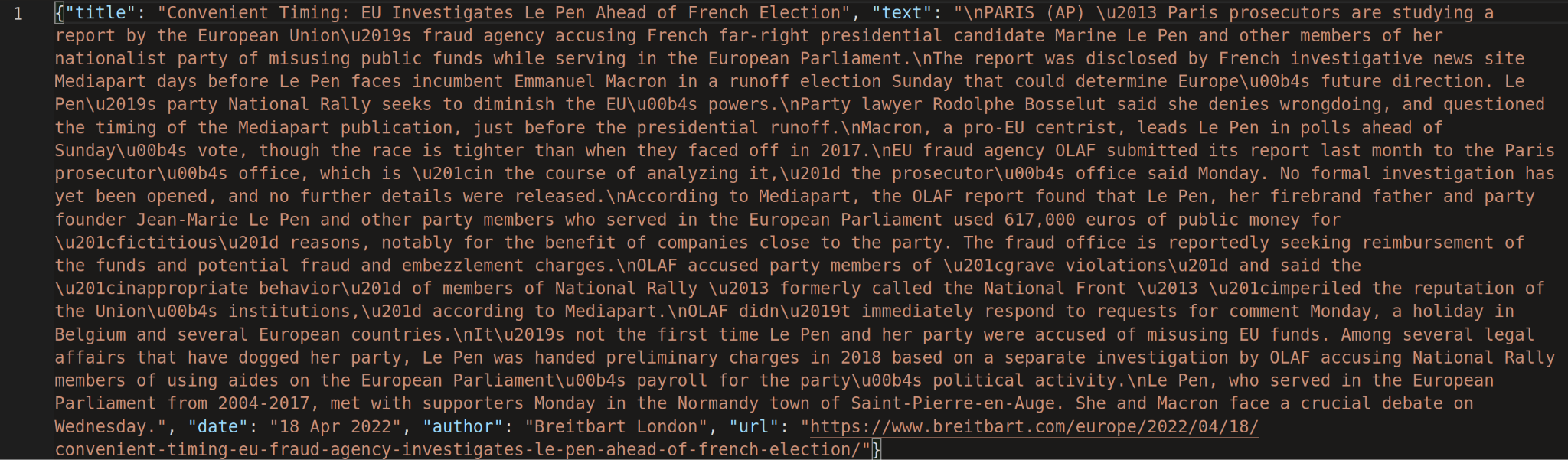
The dataset was created mostly automatically via web-scraping methodologies. Newsmax provided a simple web page with complete archives of posted content. Using this web page would provide a list of Universal Resource Locators (URL). Breitbart uses statically paginated content. Clicking ‘next page’ would change the URL and show different content. By programmatically searching the paginated results, we could obtain the list of URLs.

Fox News had no simple method for getting article URLs. Fox News provided a button that would add more articles to the current page using a JavaScript AJAX call. To retrieve a list of URLs of Fox News articles, that button had to be clicked many times, until the articles reached the beginning of the date range desired (in this case, New Year 2022). Once this button had been clicked enough times, the page’s source-HTML was saved and parsed using the Python BeautifulSoup library. Parsing this final HTML provided the list of URLs: all articles shown on that page.

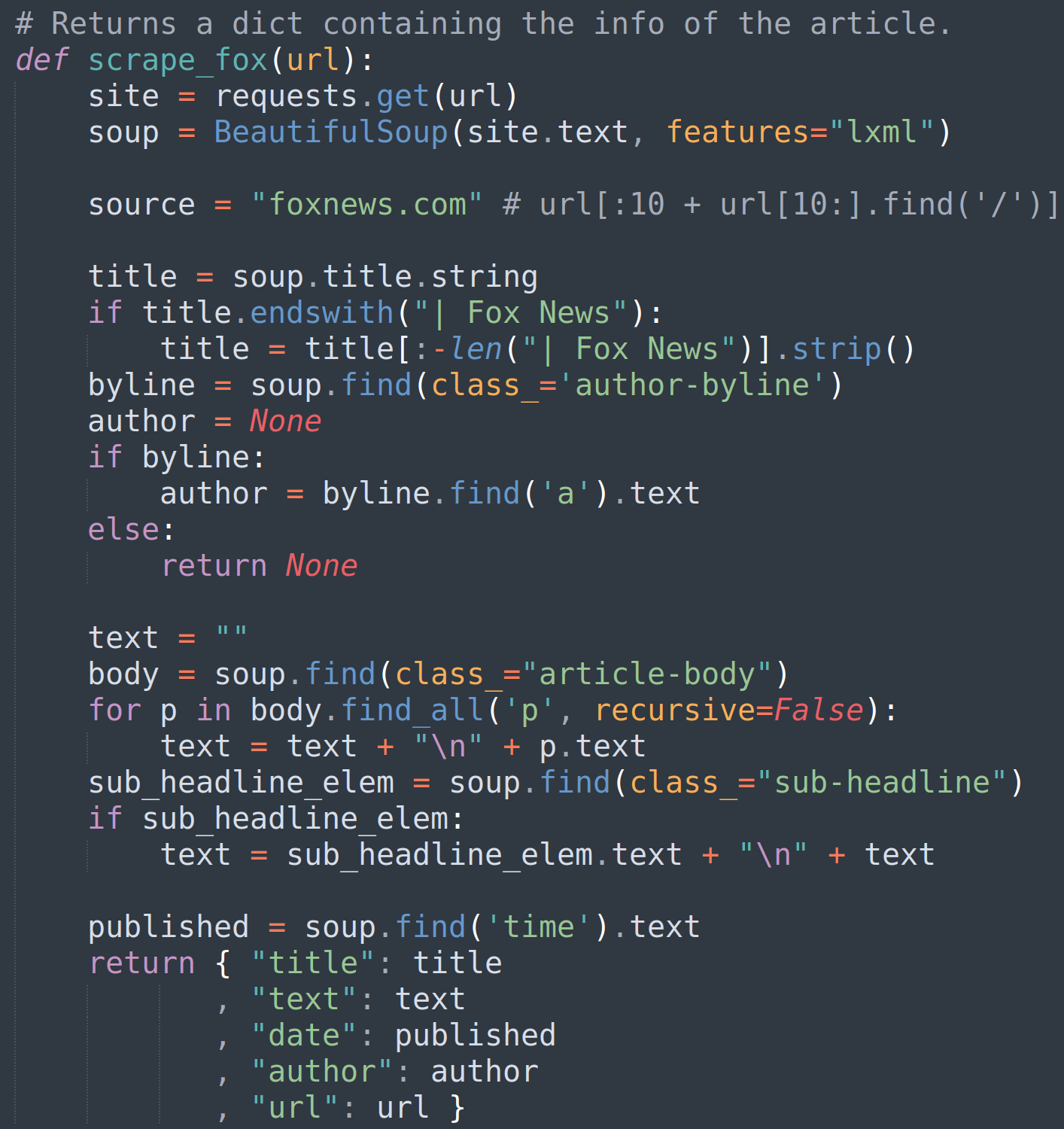




Once the lists of URLs had been made, web-scrapers were built to access those pages and then scrape them for their content. As described in the next section, all relevant data was gathered: The title (and sometimes a subtitle), article text, published date, and even the author. Each article was rendered into its own JSON object. There were three files, one for each news site, and each line contained a JSON object to be used for analysis.



Each site had a different layout, so scraping code had to be specialized. This was done by visual inspection of site source code, and then utilizing BeautifulSoup to write a scraper, parsing out the content, and saving it correctly. Article content was regularly split between many HTML tags, so care had to be taken that the entire article content was retrieved.

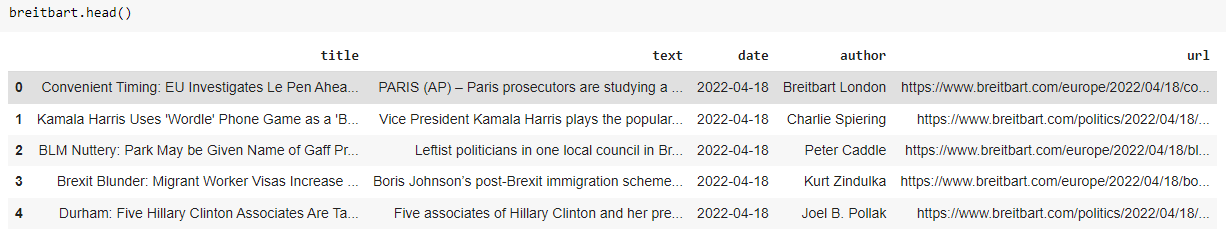


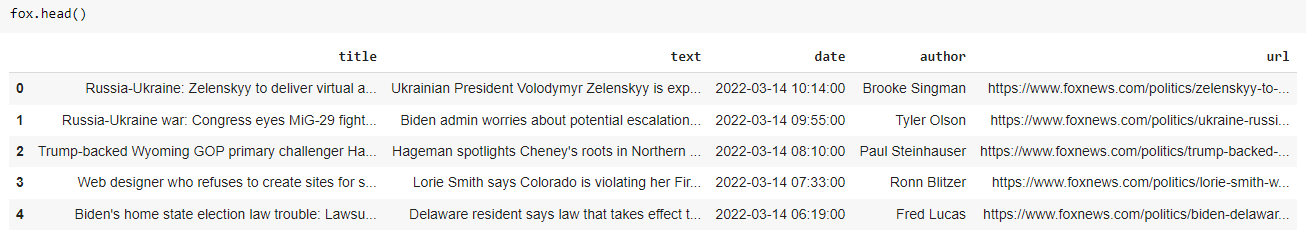
This process took many hours, as properly web scraping requires timing requests to not overload the server. A Denial of Service (DOS) attack is easy to accidentally create. However, it is also easy to spot one, and then the scraper would be IP-banned from accessing the websites. To be safe, we required waiting for 1 second after a page finished being scraped. Overall, that increased the scraping time dramatically. Scraping a page may take over 1 second, and then 1 more second must be spent waiting, evading detection from the sites.

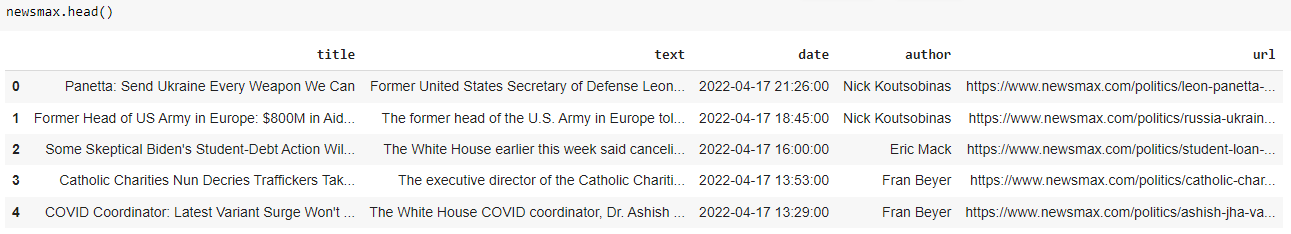
Overall, over 10 hours of compute-time were needed to gather the data. However, over 50MB of data was gathered. Months of articles by these sites.

**Dataset Description**

After finishing collecting data, we saved the collected data into three JSON files according to the news media sites. For all the news data frames, there are five columns, which are ‘title’ (titles of the news), ‘text’ (news articles), ‘date’ (date published), ‘author’ (the writers of the news), and ‘url’ (website links of the news). There are 6000 records collected for Breitbart news, 3535 records for Fox, and 8590 records for Newsmax. The data frames of three data sets are listed below:



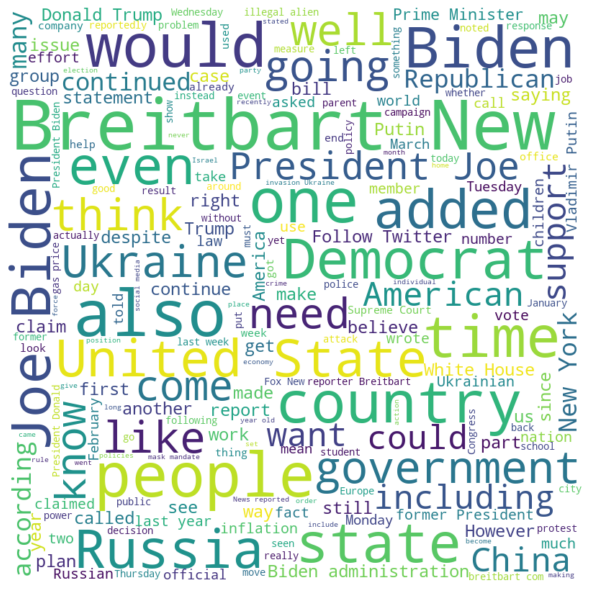


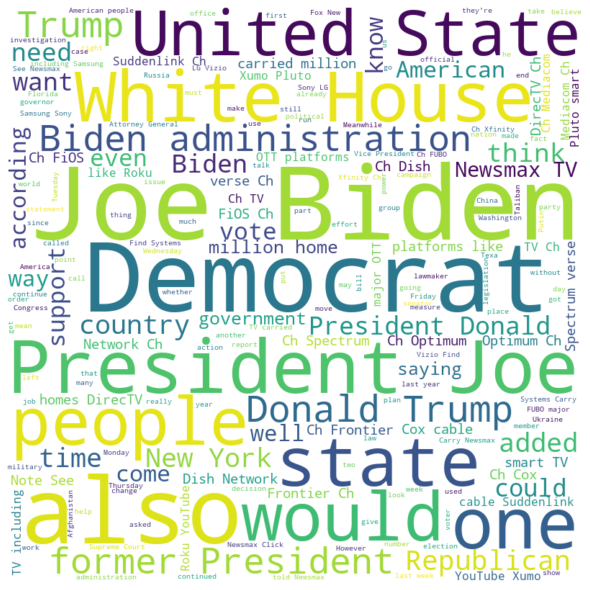


**Exploratory Data Analysis**

To understand the news content for each news site, down below, we listed word clouds for all three news sites using the words in the text column.

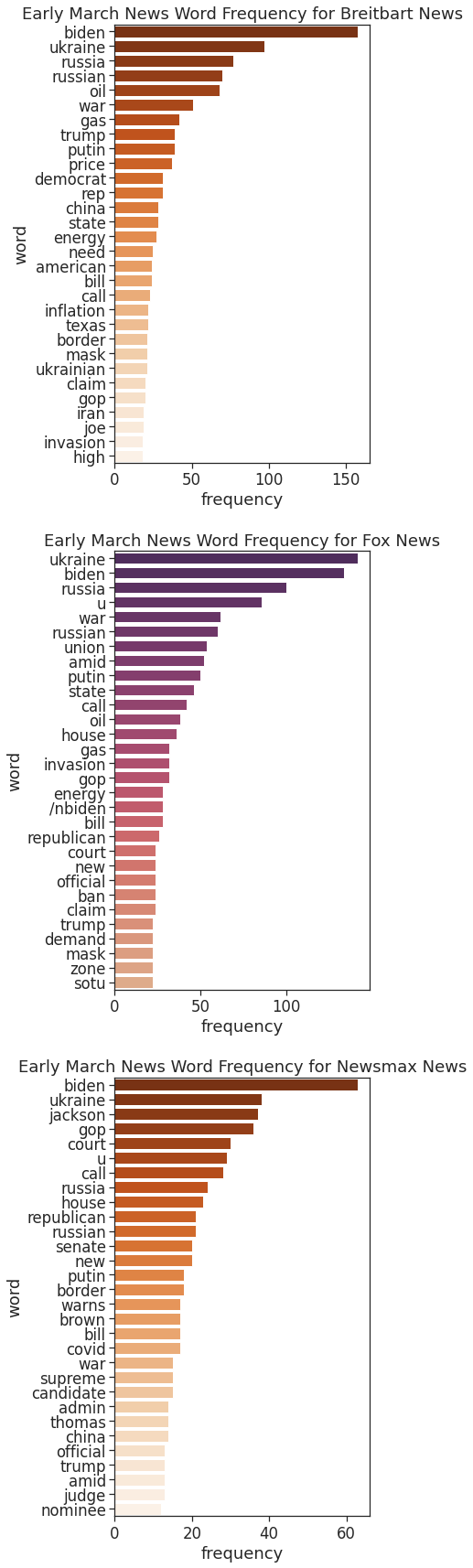
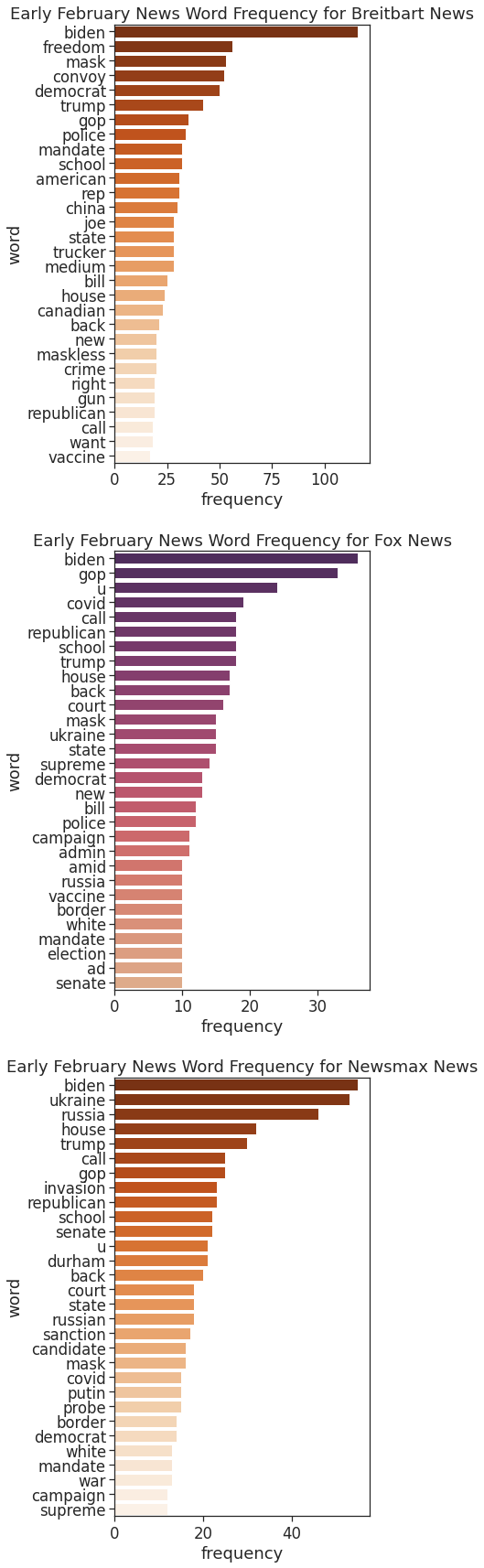
Breitbart Fox News





Newsmax

These word clouds can only give us some insights on what are the topics these news sites’ articles are trying to cover, but news articles usually are newly received or notable information. The information we obtain from the word clouds above covers a long period of time. Therefore, breaking down the news into different time periods might help us better understand each news media’s way of selecting topics to report.



We selected two time periods: February 1st to 14th and March 1st to 14th.

Early February

* Breitbart
* Fox
* Newsmax

From the result, we can see that the three news sites talk most frequently about Biden. It reflects the commonality of the topics discussed by the three news sites. What is more, We can also see the commonality between the two news sites. Like Breitbart News and Fox News, they often use words like “mask, maskless, covid”. We can roughly estimate that the topic they are talking about is related to the current epidemic situation. For Newsmax News, we can infer from the data that they are talking more about international news, because we can see some countries' names, like Ukraine and Russia. We know that war broke out between Russia and Ukraine at the end of February, but the data we collected came from the beginning of February. Compared to the other two news media companies the reports of the Newsmax seems forward-looking.

Early March

* Breitbart
* Fox
* Newsmax

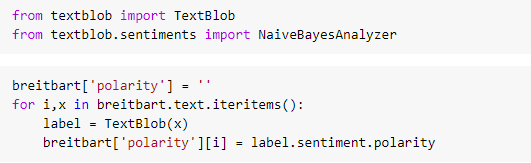
At the beginning of March, the war between Russia and Ukraine broke out in an all-around way. We can see that these three news websites are talking about the same topics, such as “Putin, war, Russian, and Ukraine”. We can also see that some words appear on these three news websites, such as oil and price, which shows that the war between Russia and Ukraine has brought great changes to the oil price. Here we take the comparative method instead of analyzing the three news websites separately. This method will be more meaningful and get more information. Through these three sets of data, we can see the characteristics of news topics, real-time, forward-looking, and commonality.

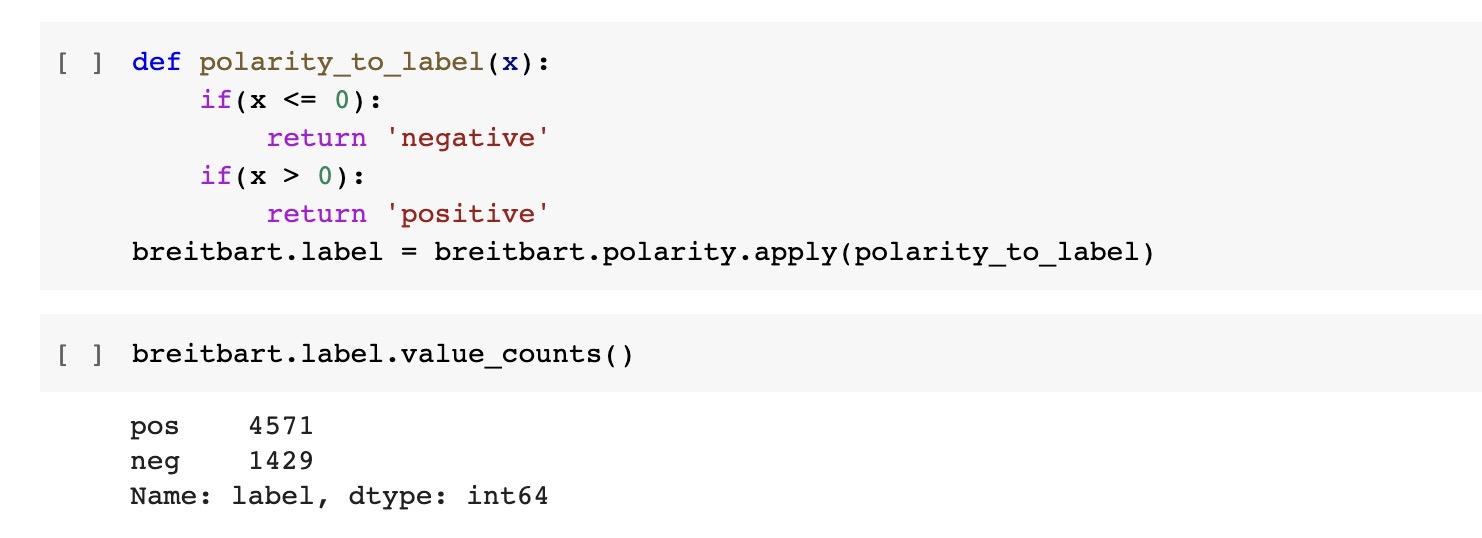
**Sentiment Analysis**

Sentiment analysis algorithms have become more efficient due to recent advances in machine learning and deep learning. The manual classification of sentiment words takes a long time. To automate the sentiment analysis process, there are two approaches that are commonly used.

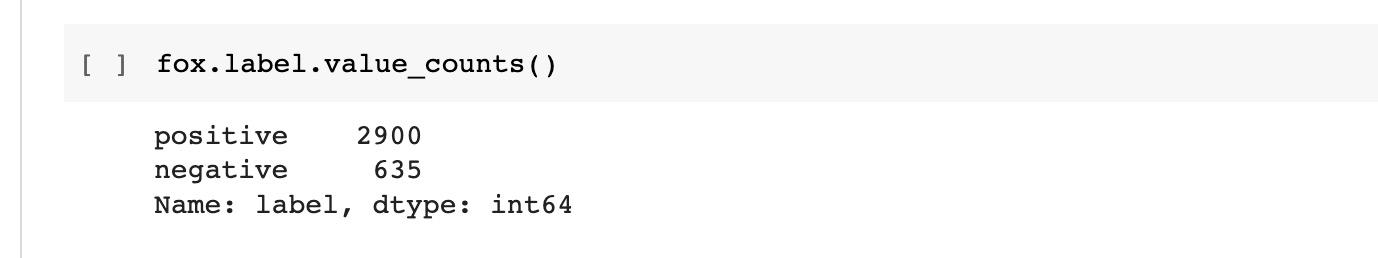
* *Rule-Based* - It matches a collection of words in a text against a word dictionary with key phrases to find polarity.
* *Automatic* - Employs machine learning methods, there is no need to preprocess data or train a classifier.

As for how to determine the sentiment polarity, we used the textblob library to perform the labeling.



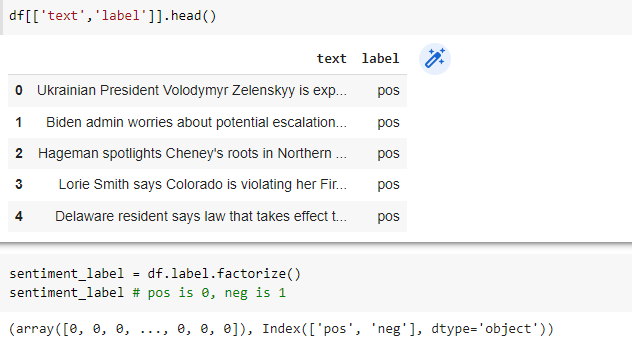




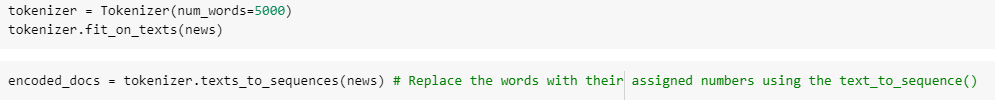


These three sets of data more clearly reflect the attitude of news websites towards using words. The positive contents outweigh the negative ones.

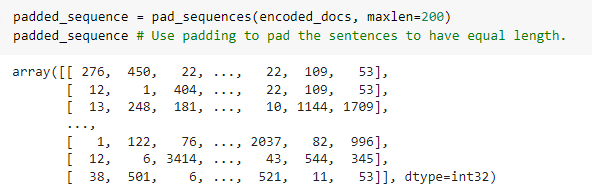
Then, we want to build a binary text classifier to classify the sentiment of text using the news data. Using the Fox News annotated dataset, we can build our own model.



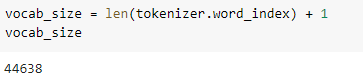
After reading the data into a data frame, we factorize the sentiment labels since the machine can only understand the numeric data.



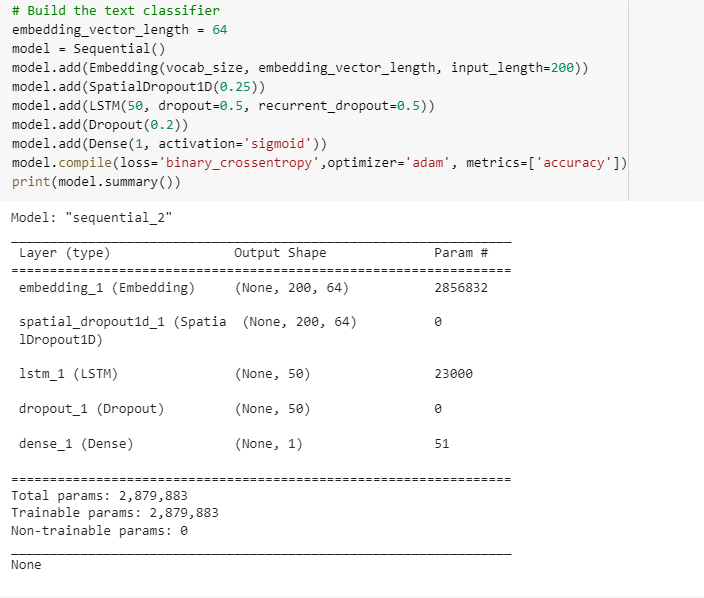
Then, we tokenized text and used texts to sequences to assign numbers to the texts.



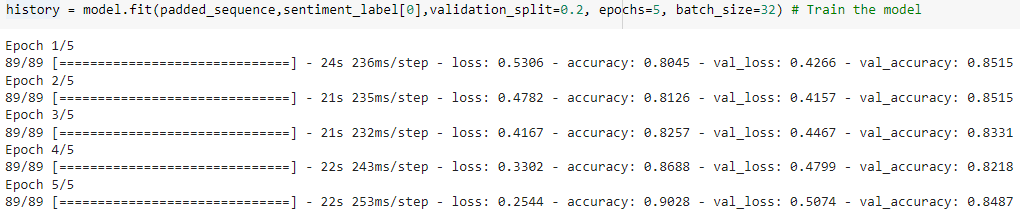
After that, we used padding to pad the sentences so that they can have equal length.



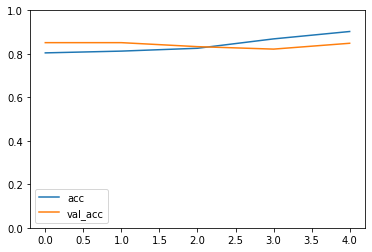
Define the vocabulary size.

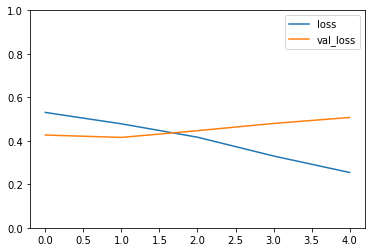


Build our text classifier.



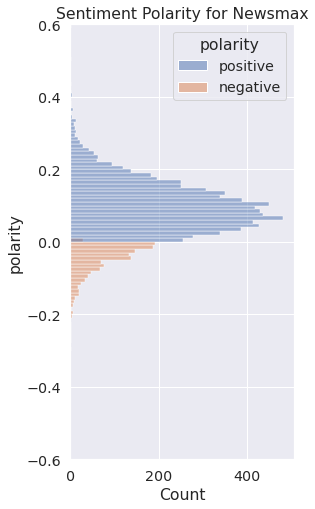
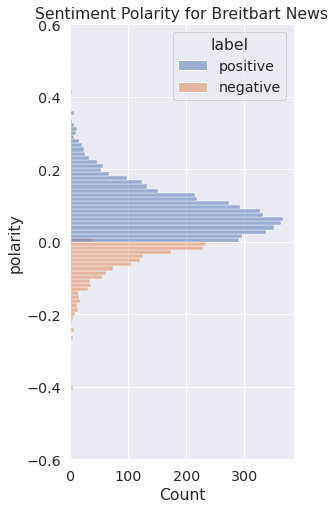
Train the sentiment analysis model.





Visualize the train (acc, loss), test (val\_acc, val\_loss) accuracies, and their losses. The overall accuracy is satisfying. As for the losses, more errors refer to higher losses, the model did not do a good job if high losses happen. Training loss is much lower than validation loss is also something we do not want to see because it means that the network might be overfitting. In our model, there might be some overfitting issues even after adjusting some hyperparameters, which means the model will not work as great when we use it on other datasets.

***Interpretation***



These are the sentiment polarities of three news sites. From the data obtained, the overall positivity in contents outweighs the negativity, this is their identical characteristic. Compared to Fox News and Newsmax, Breitbart has a slightly higher percentage of negative content. For Breitbart News and News max, the range of sentiment polarity is around 0.6. But the range of sentiment polarity for Fox News is around 0.4. It is worth mentioning again that Fox News is one of the biggest news media companies. Breitbart and Newsmax are relatively smaller companies. The wide sentiment polarity interval for Newsmax and Breitbart means that they tend to report news using more expressive words whereas Fox News is more prone to use neutral or formal words while reporting.

**Topic Modeling**

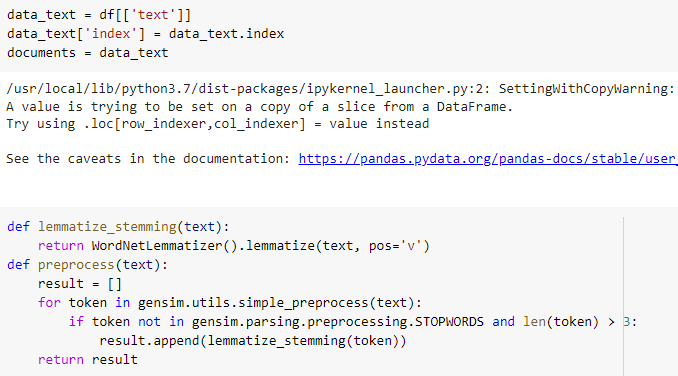
Using a probabilistic model, topic modeling discovers abstract topics that appear in a corpus of documents. It is commonly used as a text mining tool to uncover semantic structures within a body of text.

*All topic models are built on a basic assumption:*

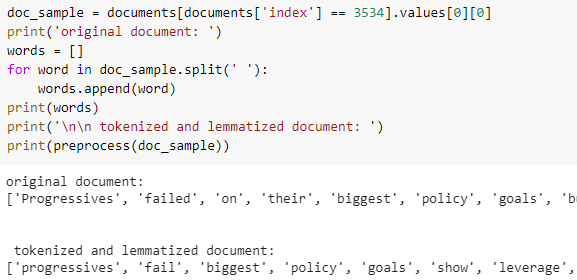
*each document is made up of a variety of topics,*

*each of which is made up of a collection of words.*

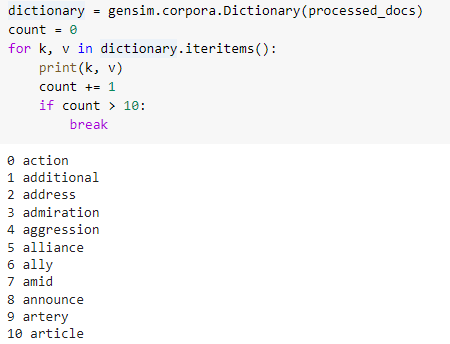
We tried to extract topics using the Latent Dirichlet Allocation (LDA) technique on fox news data. LDA is a popular Bayesian form of pLSA that's straightforward to use. For the document-topic and word-topic distributions, it uses dirichlet priors, which allows for more generalization. Since LDA may easily generalize to new documents, it is typically more successful than pLSA.



Lemmatizing text (Fox News articles)



Let’s preview a document. There are differences between original and lemmatized texts.



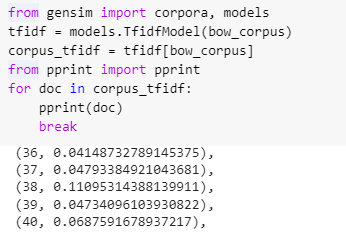
Create a dictionary



Filtering out texts that exist in less than 15 articles or more than 0.5 articles. Then, keep the first 100000 most frequent tokens.



For each article, we create a dictionary reporting how many words and times those words appear.



Apply transformation. Getting scores.



Train the LDA model using gensim.models.LdaMulticore.



***Interpretation***

Explored the words occurring in that topic and their relative weight.

The top 10 topics are listed below:

Topic: 0

Words: 0.009\*"border" + 0.008\*"administration" + 0.007\*"school" + 0.006\*"trump" + 0.006\*"covid" + 0.005\*"mask" + 0.004\*"year" + 0.004\*"texas" + 0.004\*"mandate" + 0.004\*"migrants"

Topic: 1

Words: 0.022\*"ukraine" + 0.018\*"russia" + 0.012\*"russian" + 0.009\*"nato" + 0.008\*"putin" + 0.005\*"zelenskyy" + 0.005\*"military" + 0.005\*"unite" + 0.005\*"ukrainian" + 0.005\*"force"

Topic: 2

Words: 0.021\*"russia" + 0.017\*"ukraine" + 0.011\*"russian" + 0.009\*"putin" + 0.008\*"trump" + 0.005\*"sanction" + 0.005\*"administration" + 0.005\*"invasion" + 0.004\*"continue" + 0.004\*"ukrainian"

Topic: 3

Words: 0.015\*"ukraine" + 0.009\*"putin" + 0.009\*"russia" + 0.007\*"russian" + 0.006\*"administration" + 0.005\*"trump" + 0.004\*"go" + 0.004\*"court" + 0.004\*"york" + 0.004\*"support"

Topic: 4

Words: 0.012\*"court" + 0.008\*"jackson" + 0.007\*"senate" + 0.006\*"republican" + 0.006\*"supreme" + 0.005\*"trump" + 0.005\*"campaign" + 0.004\*"judge" + 0.004\*"china" + 0.004\*"white"

Topic: 5

Words: 0.006\*"test" + 0.005\*"border" + 0.005\*"school" + 0.005\*"year" + 0.005\*"administration" + 0.005\*"mask" + 0.004\*"trump" + 0.004\*"covid" + 0.004\*"continue" + 0.003\*"go"

Topic: 6

Words: 0.014\*"vote" + 0.011\*"senate" + 0.009\*"democrats" + 0.006\*"police" + 0.006\*"republicans" + 0.005\*"right" + 0.005\*"year" + 0.005\*"republican" + 0.005\*"election" + 0.005\*"filibuster"

Topic: 7

Words: 0.008\*"trump" + 0.005\*"campaign" + 0.004\*"russia" + 0.004\*"support" + 0.004\*"right" + 0.004\*"vote" + 0.004\*"federal" + 0.004\*"texas" + 0.004\*"year" + 0.004\*"committee"

Topic: 8

Words: 0.007\*"administration" + 0.007\*"trump" + 0.006\*"ukraine" + 0.006\*"year" + 0.005\*"democrats" + 0.005\*"republican" + 0.005\*"iran" + 0.004\*"republicans" + 0.004\*"russia" + 0.004\*"nuclear"

Topic: 9

Words: 0.008\*"ukraine" + 0.008\*"trump" + 0.006\*"court" + 0.006\*"russia" + 0.005\*"republican" + 0.005\*"border" + 0.004\*"go" + 0.004\*"support" + 0.004\*"administration" + 0.004\*"democrats"

Word frequencies only tell us how often certain words appear, LDA topic modeling allocates words into topics, so we can gain a deeper understanding of the texts. For example, topic 0 has words such as “border”, “covid”, “trump”, and “mask”. It tells us that the news is about the things that happen inside America. However, in topic 1, we can see words like “russia”, “ukraine”, ”nato”, and “military”. These words informed us that this topic is about the Russia and Ukraine conflicts.

**Conclusion**

In today's world, where digital data is continually accumulating. The process of extracting useful insights from obtained data may become challenging and time-consuming. With this in mind, we concentrated on sophisticated methods such as sentiment analysis and topic modeling to accomplish such tasks automatically and efficiently. We examined sentiment analysis of news items using datasets from Newsmax, Breitbart, and Fox News. These news covers events that demonstrate emotions like positive, negative, or neutral. Sentiment analysis is used to examine human emotions found in textual data. For sentiment analysis of news stories, this project employs a lexicon-based approach. Topic modeling is also used to cluster similar words in the news collection. The extracted topics are used to gain relevant insights from the text data. We aim to showcase a representation of how news topics have changed over a period of time. There are numerous areas for exploration in sentiment analysis and topic modeling. We have focused on lexicon-based sentiment analysis in our work; however, machine learning approaches may result in more efficient solutions in future work.

**References**

* Vidiyala, R. (2021, June 8). Topic Modelling on NYT articles using Gensim,LDA. Retrieved from <https://towardsdatascience.com/topic-modelling-on-nyt-articles-using-gensim-lda-37caa2796cd9>
* Chawla, R. (2017, July 30). Topic Modeling with LDA and NMF on the ABC News Headlines dataset. Retrieved from <https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df>
* S. Taj, B. B. Shaikh and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2019, pp. 1-5, <https://doi.org/10.1109/ICOMET.2019.8673428>.
* Majumder, P. (2021, November 29). Web Scraping a News Article and performing Sentiment Analysis using NLP. Retrieved from <https://www.analyticsvidhya.com/blog/2021/11/web-scraping-a-news-article-and-performing-sentiment-analysis-using-nlp/>
* S. Taj, B. B. Shaikh and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2019, pp. 1-5, <https://doi.org/10.1109/ICOMET.2019.8673428>.
* Dataset Resources -

<https://www.breitbart.com/>

<https://www.foxnews.com/>

<https://www.newsmax.com/>