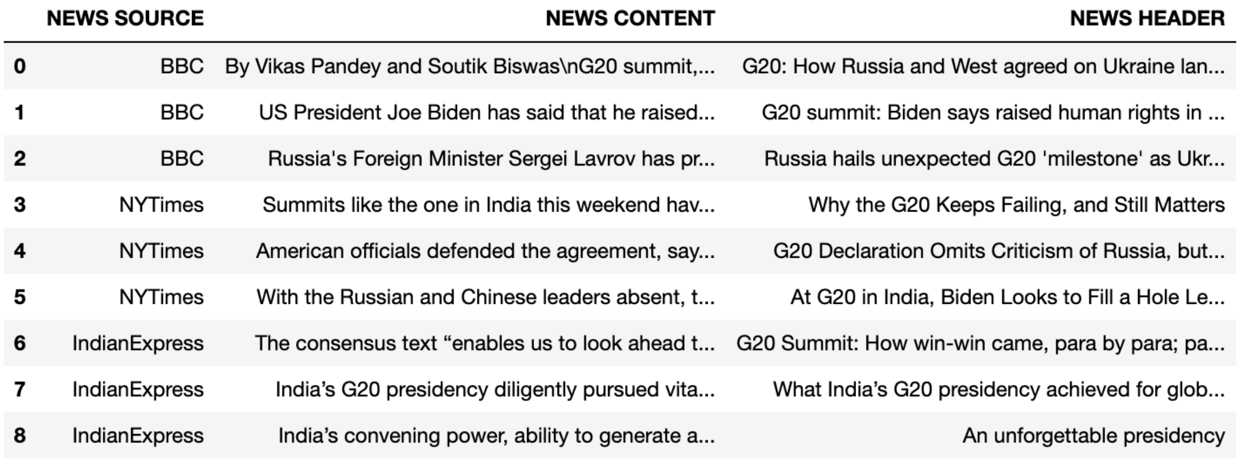
COMPARISON OF NEWS SOURCE COVERAGE OF G20 USING REPRESENTATION LEARNING - REPORT

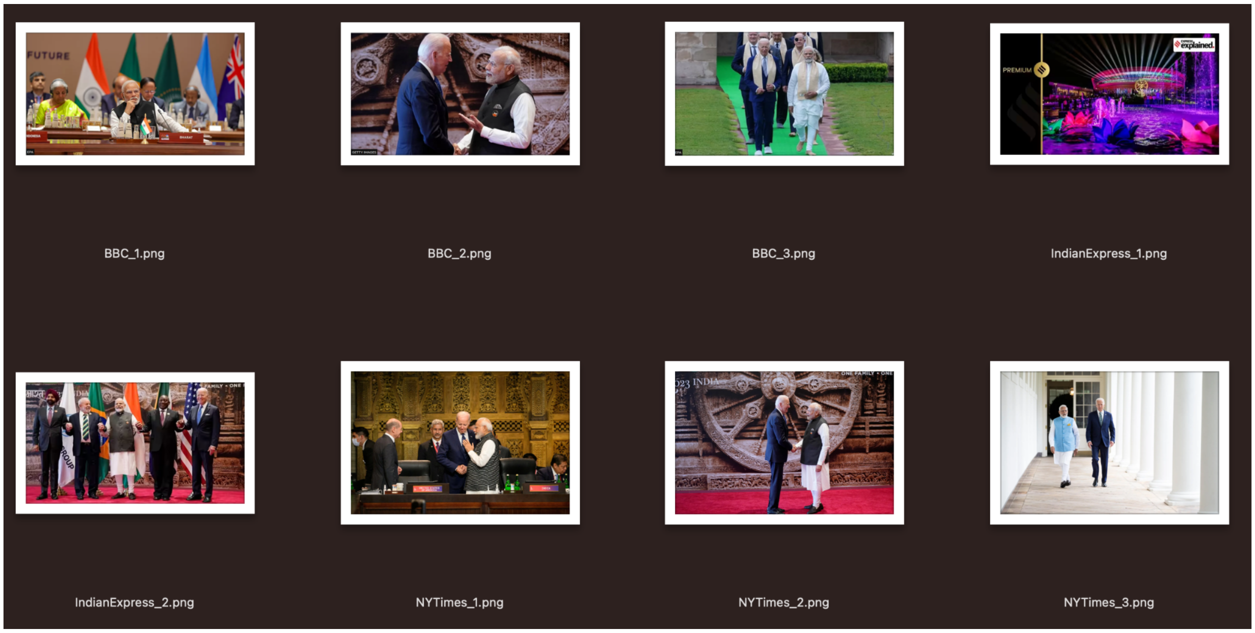
Dataset

I started by assembling a dataset of news articles from three selected sources: BBC, NYTimes, and IndianExpress. I specifically searched for news related to the G20 and identified three articles from each source, totaling nine articles. These articles were then collated into an Excel file as illustrated below:



	NEWS SOURCE	NEWS CONTENT	NEWS HEADER
0	BBC	By Vikas Pandey and Soutik Biswas\nG20 summit,...	G20: How Russia and West agreed on Ukraine lan...
1	BBC	US President Joe Biden has said that he raised...	G20 summit: Biden says raised human rights in ...
2	BBC	Russia's Foreign Minister Sergei Lavrov has pr...	Russia hails unexpected G20 'milestone' as Ukr...
3	NYTimes	Summits like the one in India this weekend hav...	Why the G20 Keeps Failing, and Still Matters
4	NYTimes	American officials defended the agreement, say...	G20 Declaration Omits Criticism of Russia, but...
5	NYTimes	With the Russian and Chinese leaders absent, t...	At G20 in India, Biden Looks to Fill a Hole Le...
6	IndianExpress	The consensus text “enables us to look ahead t...	G20 Summit: How win-win came, para by para; pa...
7	IndianExpress	India’s G20 presidency diligently pursued vita...	What India’s G20 presidency achieved for glob...
8	IndianExpress	India’s convening power, ability to generate a...	An unforgettable presidency

Additionally, for each news article I chose the representative image and saved that as well:



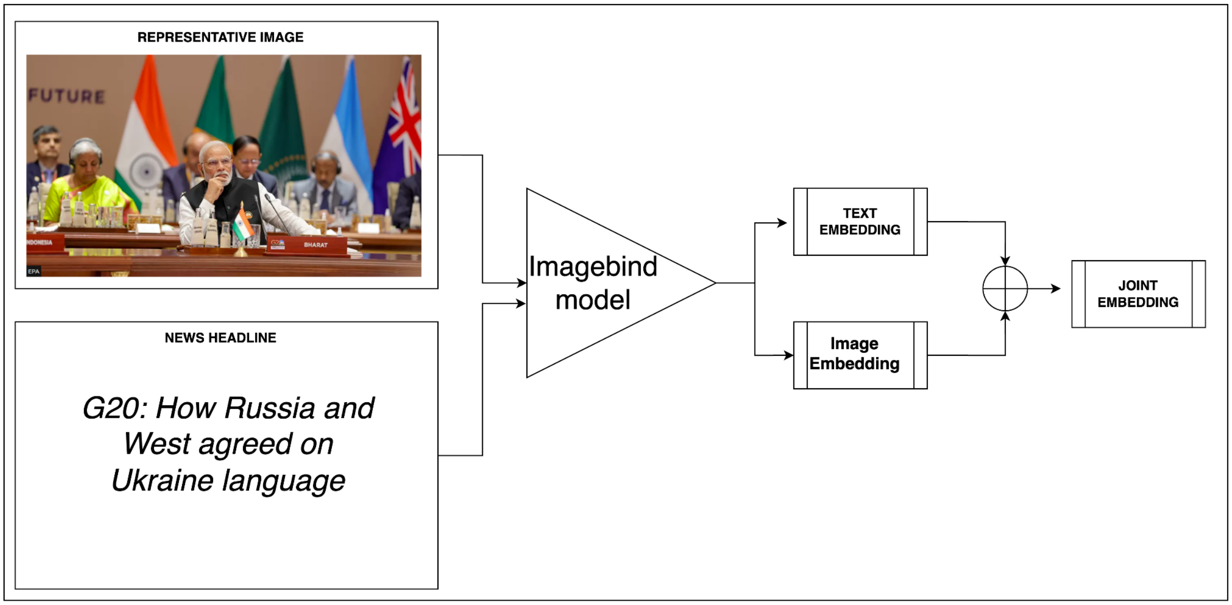
To ensure consistency, each article was paired with only one image—the one located directly beneath its headline. As a result, every news piece is represented by an image, its headline, and the article content.

ImageBind Multimodal Embeddings

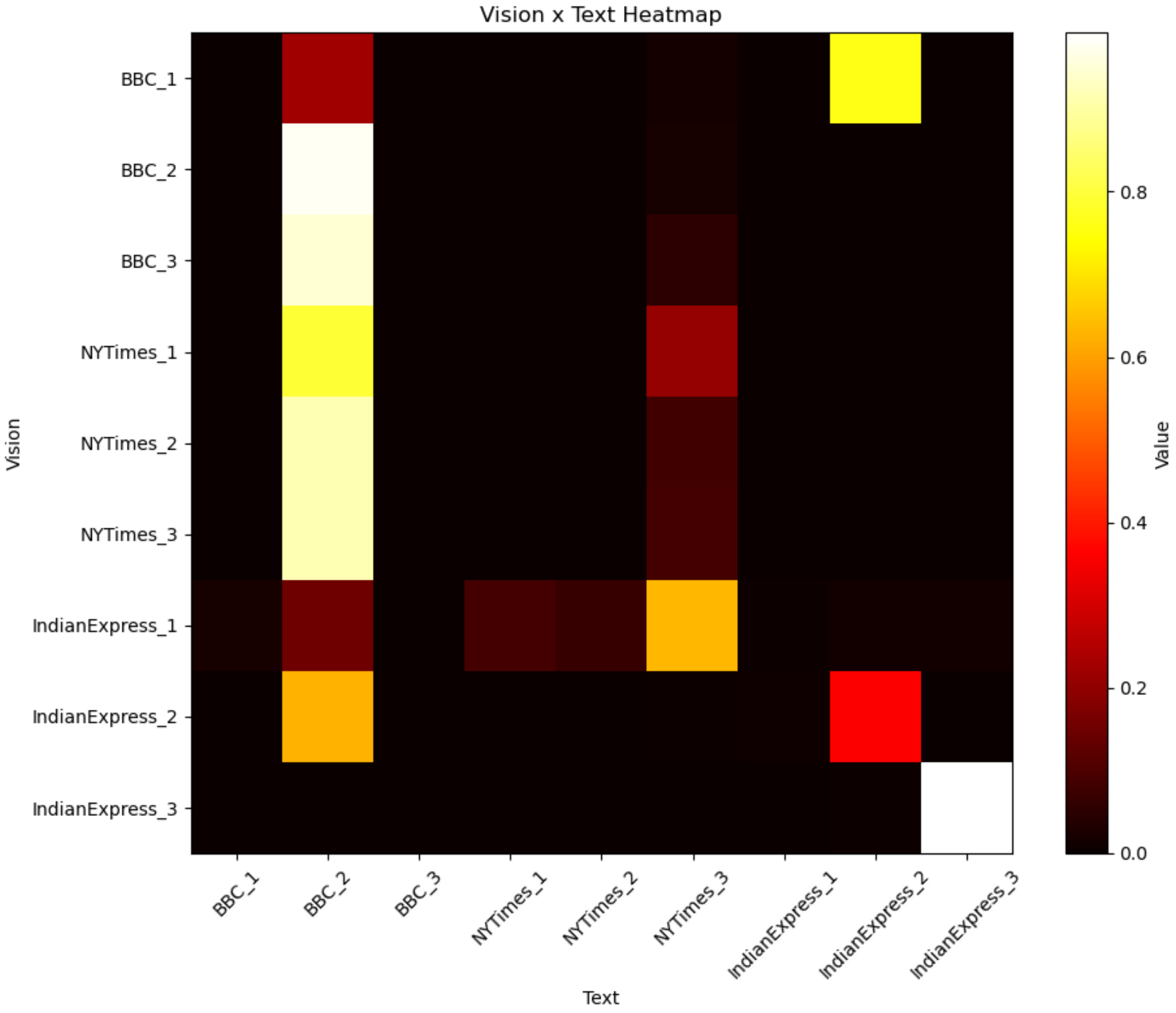
To better represent both the article's image and content, I explored the newly introduced ImageBind embeddings, a multimodal embedding method that captures text, image, audio, and other modalities from a unified embedding space. This approach promises improved coherence and facilitates detecting cross-modal similarities. However, during my experimentation, I realized that ImageBind employs a transformer model for text encoding, which has a limit of 77 tokens. Given this limitation, I opted to use ImageBind embeddings exclusively for the combination of the article's representative image and its headline.

These generated embeddings now serve as a tool to discern differences and similarities among various sources and articles. This can be achieved by performing operations like taking a dot product followed by a softmax function or calculating the cosine similarity. I will delve deeper into these findings in subsequent analyses.

Methodology for extracting text, image and joint embeddings using the ImageBind model



Analysing Visual-Text Relationship



Code Explanation:

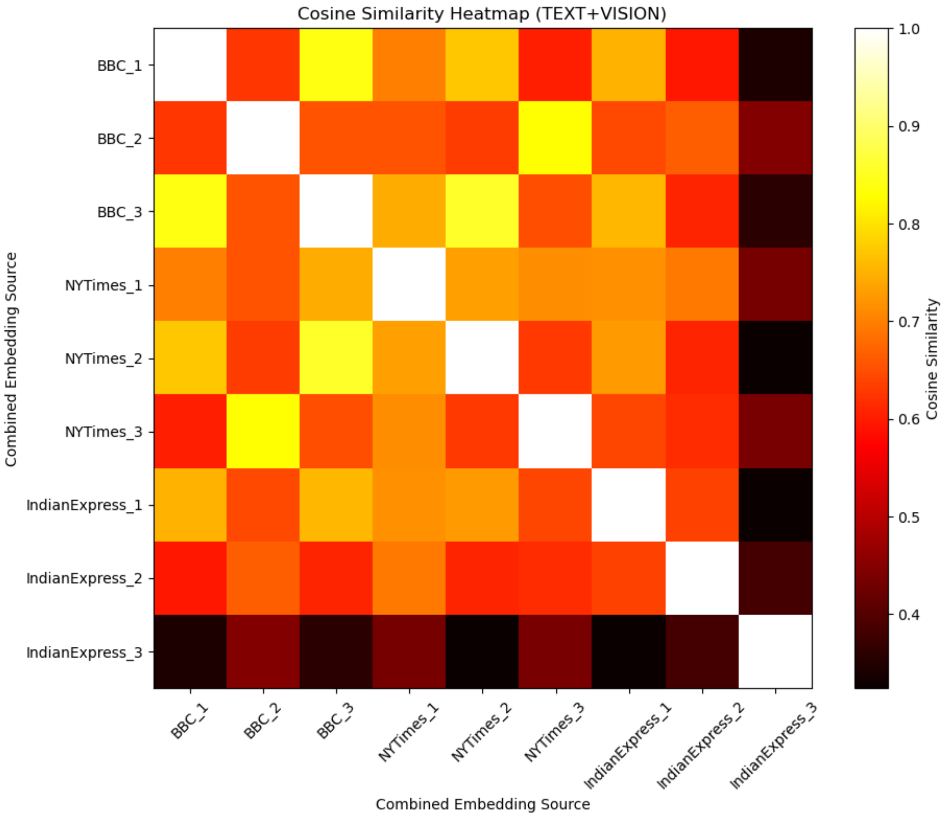
The script aims to represent the correlation between vision (image) embeddings and text embeddings of various news items in a visual format, namely a heatmap. Here's a breakdown:

- **Matrix Formation:** The script has a matrix where each cell value signifies the correlation between vision and text embeddings. If you look at row i and column j, the value at this intersection (i.e., (i, j)) represents the relationship between the (i)-th image and the (j)-th text.

- **Labelling:** The identifiers, such as 'BBC_1', 'BBC_2', etc., help distinguish the source and a unique number to each news item. For example, 'BBC_1' could represent the first news item from BBC.

Heatmap Analysis:
The similarities between visuals and headlines are generally minimal. However, the second BBC headline, "G20 summit: Biden says raised human rights in India with Modi" (referred to as ‘BBC_2’), displays notable similarity with many other articles, especially those from BBC and NYTimes sources. This observation is anticipated given that the entities mentioned in the headline are featured in the images. Likewise, the ‘NYTimes_3’ headline bears resemblance to images from other articles, but is particularly coherent with its own images, suggesting a consistent image trend. The image from 'bbc_1', which features PM Modi, aligns with the headline of 'IndianExpress_2', titled "What India’s G20 presidency achieved for global health". The 'IndianExpress_2' headline and its image also share significant similarity. It appears that images of notable figures or groups are more likely to resonate with headlines addressing global health issues.

Analysing (Visual+Text)x(Visual+Text) Relationship



Code overview

The code utilizes a heatmap to vividly display the cosine similarity between combined embeddings of news items. Distinct from the previous heatmap, this one focuses on similarities based on a joint embedding of both the news image and headline. This methodology ensures a more holistic assessment of similarity, considering the mutual influence of both visual and textual components. Each cell in the matrix (i, j) indicates the cosine similarity between the i-th and j-th combined embeddings. To further contextualize the data, identifiers such as 'BBC_1', 'BBC_2', etc., are used, representing the news source coupled with a unique article identifier.

Interpreting the Heatmap:

BBC & NYTimes (Western News Sources):

BBC and NYTimes: The cosine similarity values between the BBC and NYTimes articles tend to be on the higher side. For instance, when comparing 'BBC_1' with 'NYTimes_1' or 'BBC_3' with 'NYTimes_2', there's a discernible affinity. This points to the possibility that Western news outlets like BBC and NYTimes may have overlapping content or similar reporting styles, especially when considering both their images and headlines.

Within BBC and NYTimes: There's also considerable similarity within individual sources, evident from comparisons like 'BBC_1' with 'BBC_2' or 'NYTimes_1' with 'NYTimes_3'. This suggests a level of content consistency in these Western media houses.

IndianExpress:

With Western Sources: When comparing IndianExpress to its Western counterparts, such as BBC and NYTimes, the cosine similarities tend to be lower. For instance, the similarity values between 'IndianExpress_1' and both 'BBC_1' and 'NYTimes_1' are markedly reduced. This suggests that the amalgamated textual and visual content from IndianExpress often stands distinct from Western news narratives. Yet, it's crucial to highlight that these patterns are not overwhelmingly pronounced; overall, IndianExpress articles exhibit diminished similarities with other articles, irrespective of the source.

Conclusively, while Western news sources like BBC and NYTimes demonstrate closer affinity amongst themselves, IndianExpress charts a more distinct path, especially when juxtaposed with Western media.

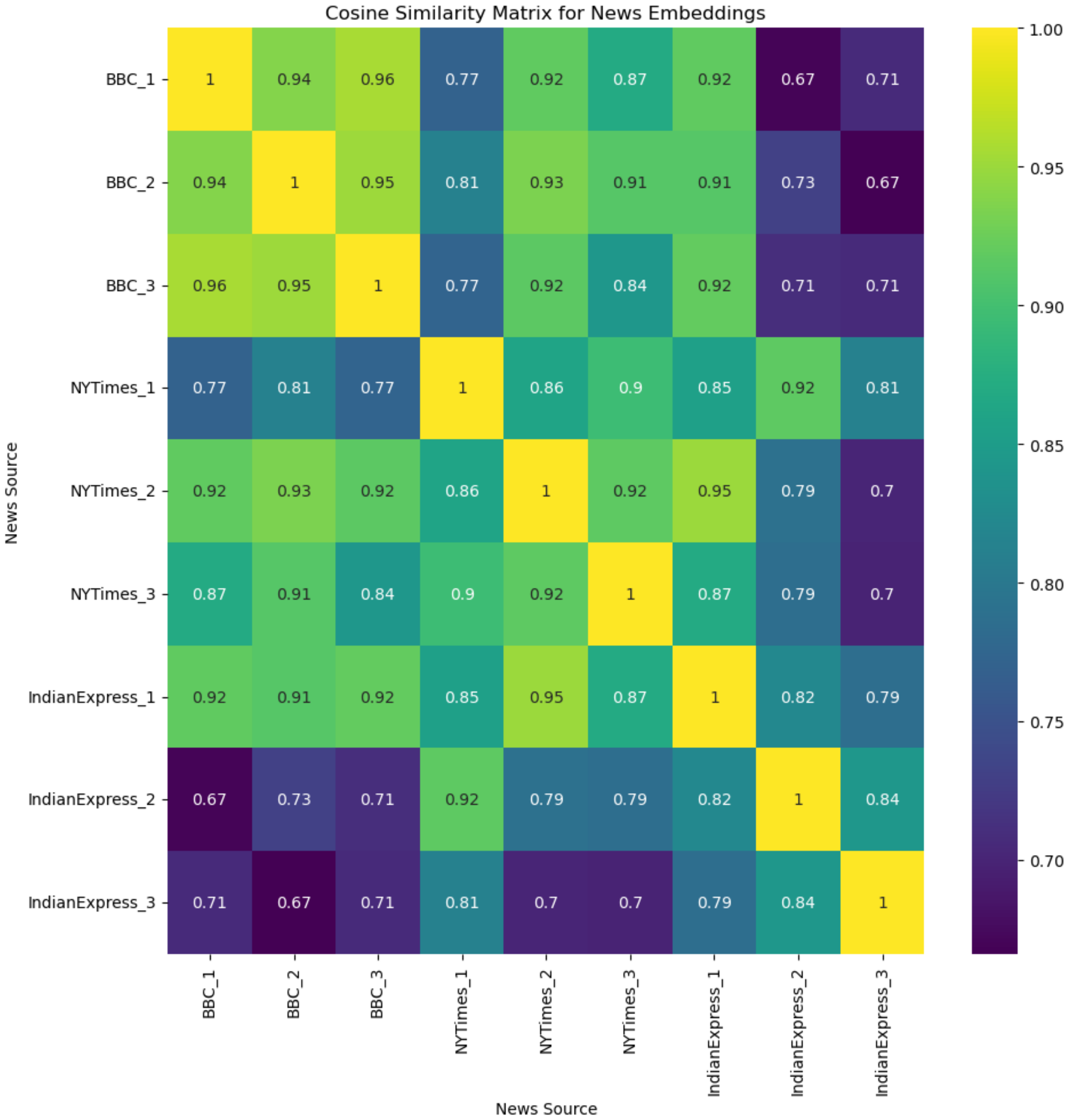
Analysing Article Content Relationship

In order to compare the actual article content I wanted to find an embedding that has been trained on larger text and especially text that is written in the style of news articles so that the nuances in the news articles can be captured. Therefore, I utilised the BERT-mini model trained for classification task on the AG news dataset. Which is a compilation of 1 million news articles.

Code explanation:

The code imports necessary libraries and loads textual content from 'text_content.pkl'. Using the Huggingface Transformers library, it initializes a pre-trained BERT model for news category classification. For each article in the dataset, the script tokenizes the text, predicts its news category, calculates associated probabilities, and extracts the embeddings from the model's last hidden layer. These embeddings are saved in 'news_embeddings.pkl' for future use. Throughout, it also prints the original text, its predicted category, and the probabilities for each class.

Heatmap explanation:


Cosine Similarity Matrix for News Embeddings

	BBC_1	BBC_2	BBC_3	NYTimes_1	NYTimes_2	NYTimes_3	IndianExpress_1	IndianExpress_2	IndianExpress_3
BBC_1	1	0.94	0.96	0.77	0.92	0.87	0.92	0.67	0.71
BBC_2	0.94	1	0.95	0.81	0.93	0.91	0.91	0.73	0.67
BBC_3	0.96	0.95	1	0.77	0.92	0.84	0.92	0.71	0.71
NYTimes_1	0.77	0.81	0.77	1	0.86	0.9	0.85	0.92	0.81
NYTimes_2	0.92	0.93	0.92	0.86	1	0.92	0.95	0.79	0.7
NYTimes_3	0.87	0.91	0.84	0.9	0.92	1	0.87	0.79	0.7
IndianExpress_1	0.92	0.91	0.92	0.85	0.95	0.87	1	0.82	0.79
IndianExpress_2	0.67	0.73	0.71	0.92	0.79	0.79	0.82	1	0.84
IndianExpress_3	0.71	0.67	0.71	0.81	0.7	0.7	0.79	0.84	1

The provided heatmap matrix illustrates the cosine similarities among nine news article contents from three sources: BBC, NYTimes, and IndianExpress. A quick analysis of the matrix shows:

High Internal Consistency within Western Media: The articles from BBC ('BBC_1', 'BBC_2', 'BBC_3') and NYTimes ('NYTimes_1', 'NYTimes_2', 'NYTimes_3') share very high similarity values among themselves, often exceeding 0.8. This suggests a significant degree of content overlap or shared topical focus within each of these Western sources and even across them.

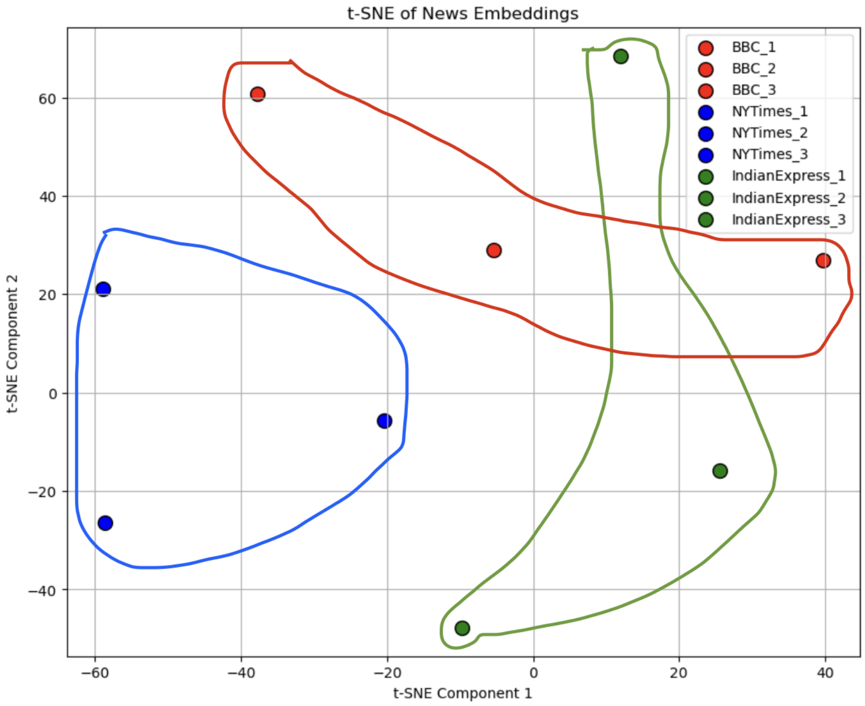
Distinctiveness of IndianExpress: When we look at the IndianExpress articles' similarity with Western articles, the values generally fall below 0.8, often even below 0.75. This implies that the content of IndianExpress is often distinct from the Western media narratives. For instance, 'IndianExpress_2' has a similarity of 0.67 with 'BBC_1', showing a notable divergence.

Internal Consistency within IndianExpress: However, when comparing the IndianExpress articles amongst themselves ('IndianExpress_1', 'IndianExpress_2', 'IndianExpress_3'), we observe a stronger similarity, often above 0.79. 'IndianExpress_2' and 'IndianExpress_3', for example, have a similarity of 0.84, indicating that while they may diverge from Western paradigms, they maintain consistency in their own narratives.

In summary, while Western media sources like BBC and NYTimes exhibit high mutual similarities, suggesting shared narratives, the IndianExpress presents a more unique perspective. However, within its own articles, the IndianExpress maintains a consistent narrative thread.

Creating TSNE plot from news content embeddings

A t-SNE plot of news content embeddings provides a visual representation of the high-dimensional data in a 2D space, effectively capturing the nuances and relationships between different news sources. By projecting these embeddings, we can discern patterns, groupings, or disparities, enabling a quick and intuitive understanding of the similarities or dissimilarities between news sources. Such visual insights can be invaluable in asserting the uniqueness or overlap of content paradigms among different media outlets.



The t-SNE plot reveals distinct clustering patterns, particularly evident in the NYTimes cluster. While there's a semblance of overlap between the BBC and IndianExpress clusters, the pronounced separations underscore the disparities in G20 event coverage by Western and Indian news outlets.

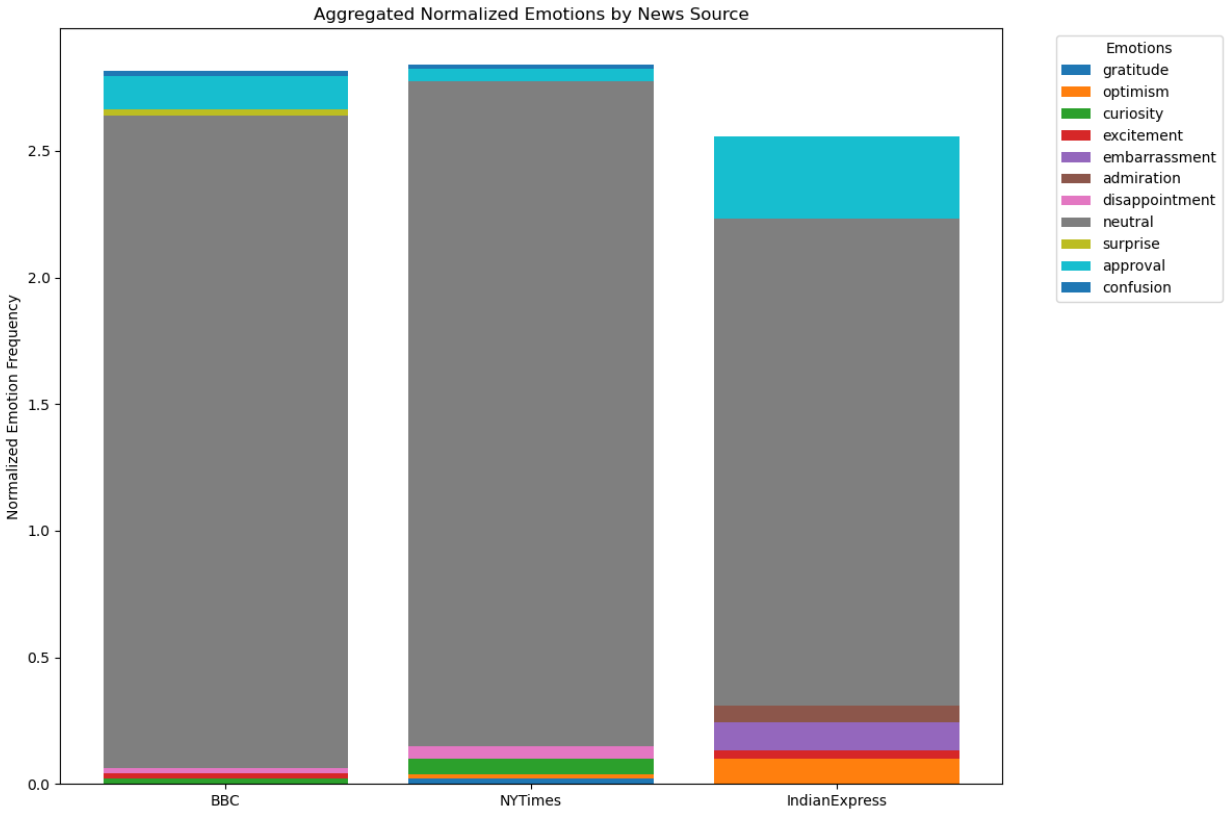
The assessments thus far offer compelling evidence of this discernability. Moving forward, my focus will pivot to dissecting the nuanced differences in coverage, encompassing sentiment, emotional undertones, and other textual metrics, as detailed in the following sections.

Comparison of emotions in the text

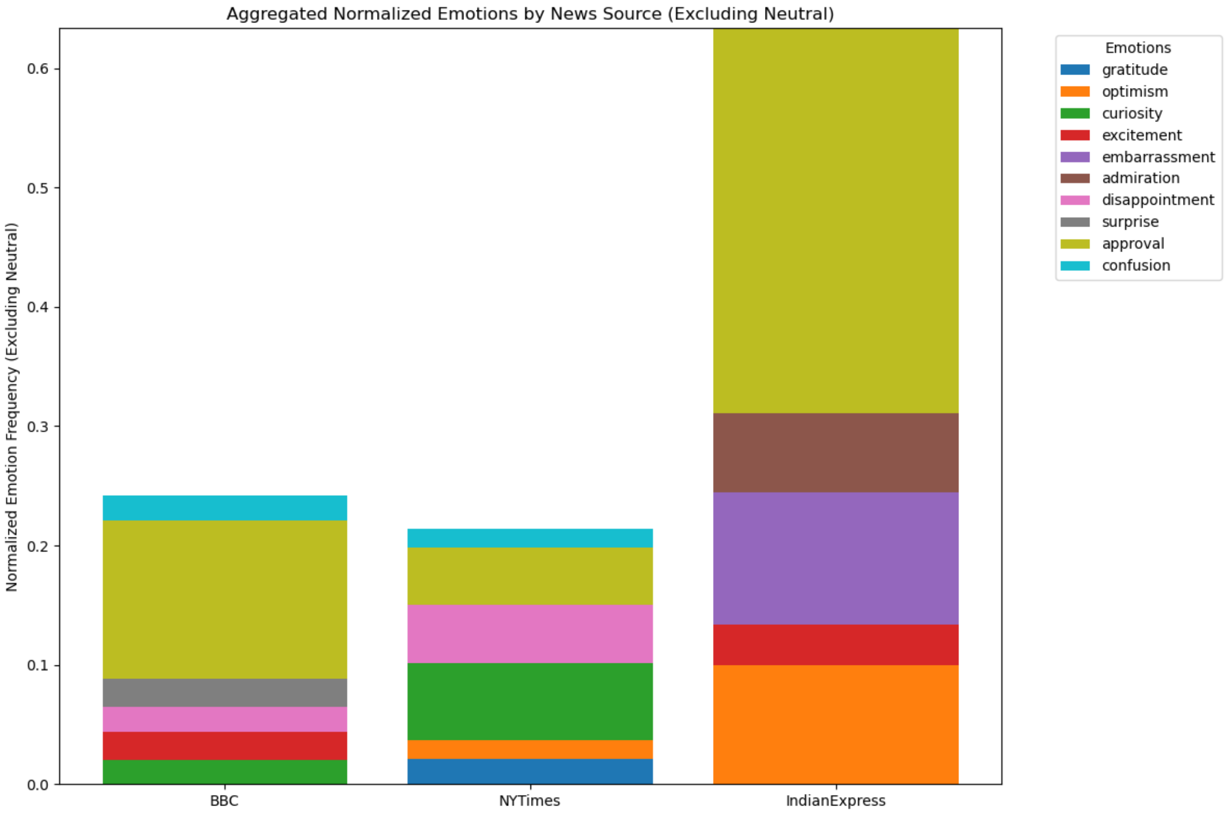
Code overview

The code uses the `roberta-base-go_emotions` model, trained on the GoEmotions dataset, to perform emotion analysis on news articles. The GoEmotions dataset comprises 58k meticulously annotated Reddit comments, labeled across 27 emotion categories or as Neutral. Given that the dataset is optimized for shorter text segments like comments, the code breaks down each article into individual lines using the NLTK library. It then assesses the emotional sentiment of each line. Finally, detected emotions from all lines of an article are aggregated to yield a consolidated emotional footprint for that article, normalized by the number of lines.

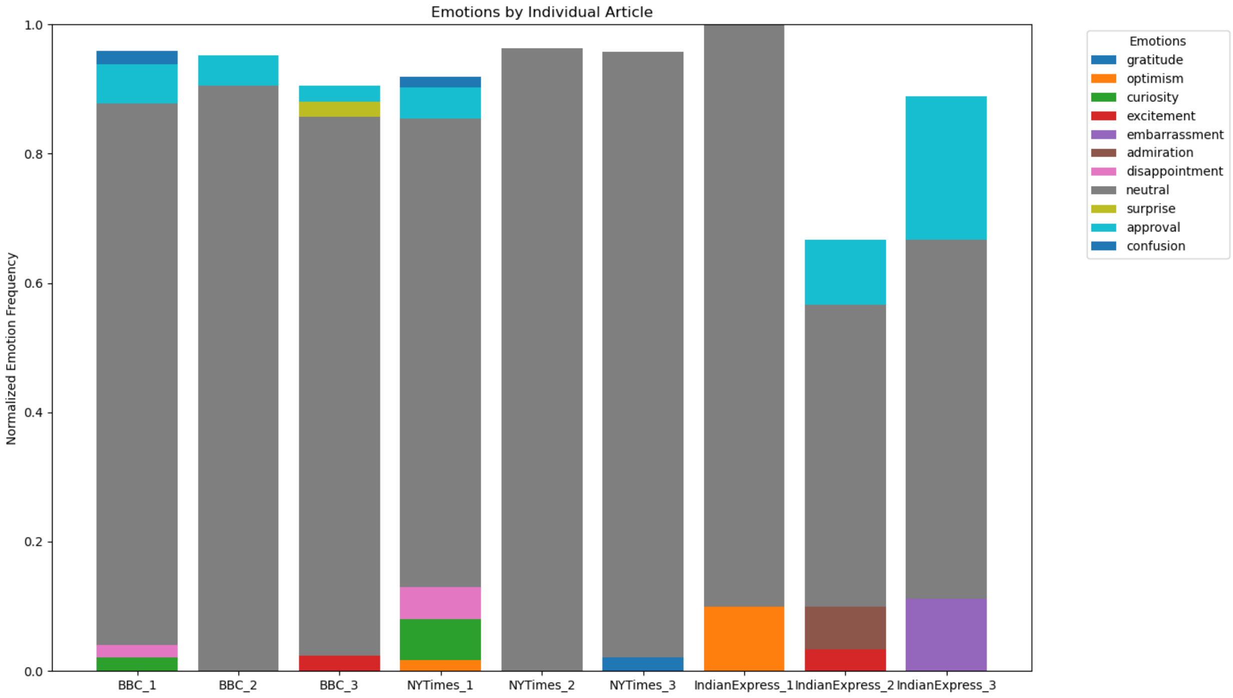
PLOT - 1: Emotions for each news source (aggregated)



PLOT - 2: Emotions for each news source (aggregated) with neutral removed



PLOT - 3: Emotions for each news article

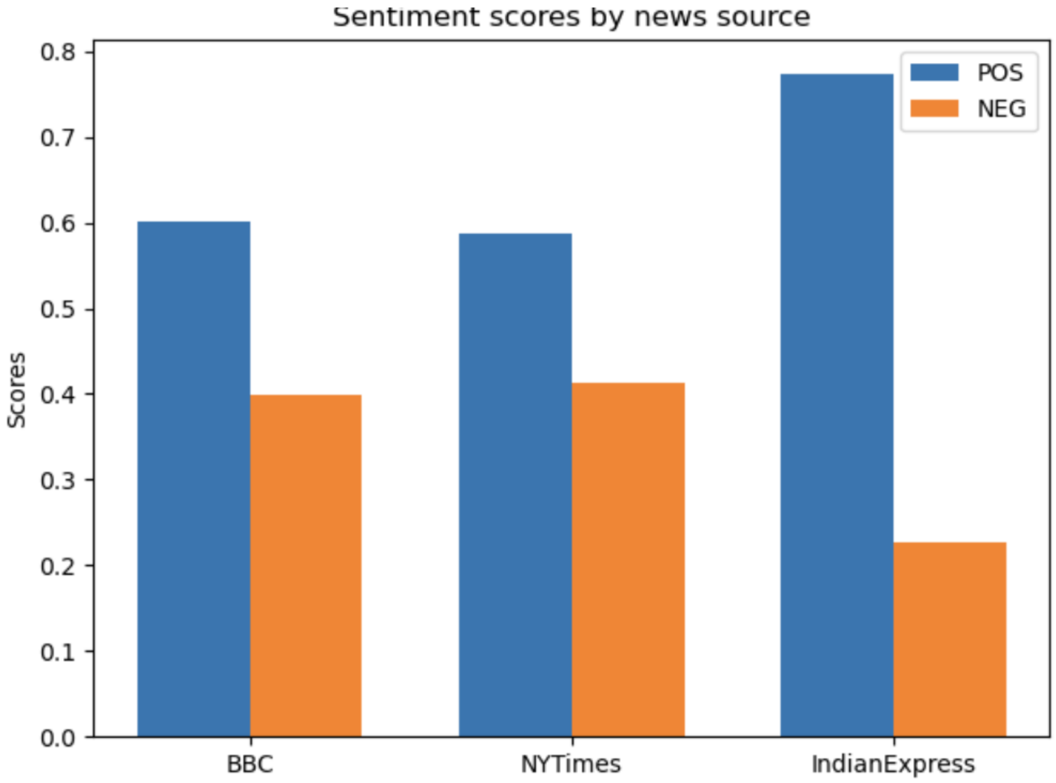


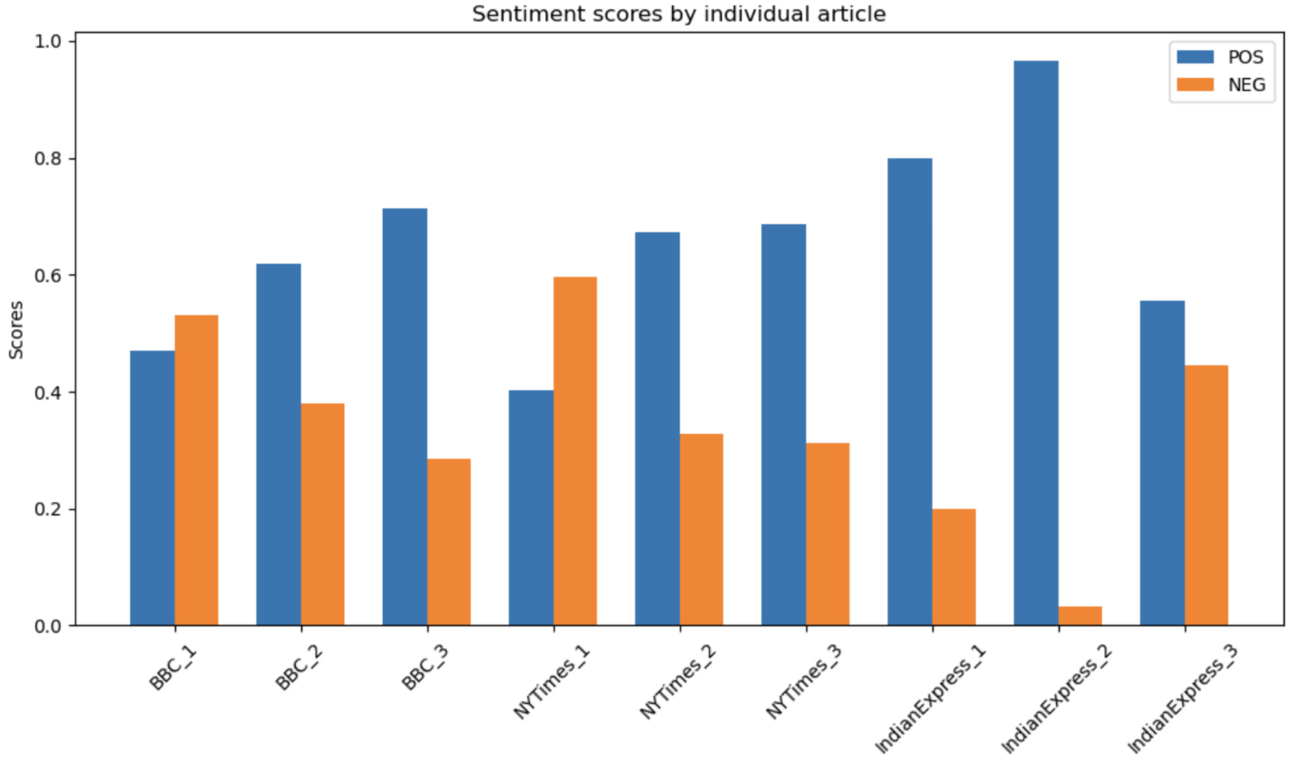
Based on the emotional analysis, Western news outlets like BBC and NYTimes predominantly exhibit a 'neutral' tone in their articles, as illustrated in PLOT-1. Excluding the 'neutral' emotion in PLOT-2, these Western publications manifest secondary emotions such as 'approval' and 'curiosity'. Conversely, the IndianExpress articles showcase a diverse range of emotions. Notably, sentiments like 'optimism', 'gratitude', and 'admiration' are more pronounced in the IndianExpress content. PLOT-3 further underscores 'approval' as a consistent emotion in IndianExpress, with the inclusion of 'admiration', indicating a positive approach to their stories. Absence of emotions like 'confusion', 'disappointment', and 'curiosity' from IndianExpress's coverage is evident. The emergence of 'embarrassment' is noteworthy, though its exact context demands a detailed examination of the content. In summary, while Western media subtly casts a skeptical light on the G20 event, maintaining a measured narrative, IndianExpress offers a more affirmative and positive perspective in its dispatches.

Comparison of sentiment in the text

Code overview:

The code initiates by loading the articles' content and dividing them into individual lines using the sent_tokenize function from the NLTK library. The sentiment analysis employs the Distil-BERT model which is trained on the SST-2 dataset, a widely used benchmark for sentence-level sentiment classification. After analyzing each line from the articles, the sentiments are aggregated to deduce a comprehensive sentiment score for each article. This model specifically classifies sentiments into either "POSITIVE" or "NEGATIVE", without a neutral designation.


Sentiment scores by news source


Sentiment scores by individual article

Analysis of Sentiment Results:
Comparing sentiments across articles indicates a divergence in tonalities between Western and Indian news sources. BBC articles demonstrate a fluctuation between positive and negative sentiments, with 'BBC_1' showing a slight inclination towards the negative spectrum. Meanwhile, the NYTimes set, although varied, has 'NYTimes_1' distinctly presenting a negative hue. Conversely, IndianExpress articles, especially 'IndianExpress_2', exude a compelling positive sentiment, showcasing an overarching optimistic narrative.

On the whole, while articles from BBC and NYTimes manifest a balanced amalgamation of positive and negative sentiments, IndianExpress stands out with its pronounced positive tilt. This suggests that while Western media incorporates moments of scrutiny and skepticism, IndianExpress offers a brighter perspective in its reporting.

Comparison of topics covered

Top topics in BBC:
India, summit, Modi, Russia, Ukraine

Top topics in IndianExpress:
India, health, global, digital, challenge

Top topics in NYTimes:
Biden, country, Modi, leader, India

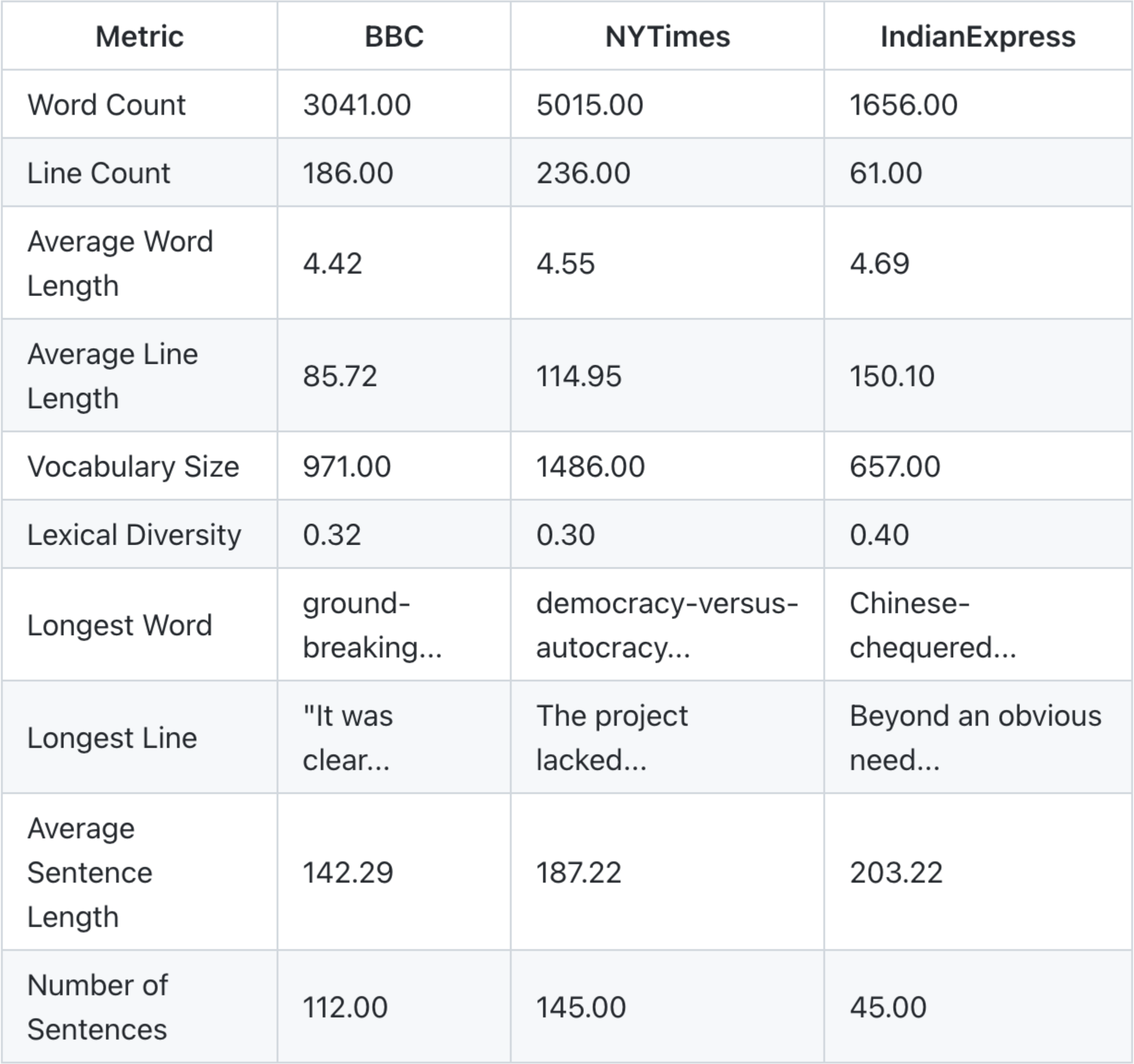
Code overview:

The script employs the Latent Dirichlet Allocation (LDA) algorithm for topic modeling. It begins by loading news articles from an Excel file and aggregating them by source. These articles are then preprocessed through tokenization, removal of stopwords, lemmatization, and filtering out of short words. Using this refined data, a dictionary of unique words and a bag-of-words representation for each article are generated. The LDA model is then trained on this corpus, iterating 15 times to discern specific topics within the aggregated content. For each news source, the most relevant topic and its associated key terms are displayed.

Analysis of results:
Upon examining the results, we observe that the LDA model identifies distinctive primary topics for each news outlet. BBC's coverage appears to pivot around international politics and diplomacy, especially emphasizing India's role or interactions, as indicated by terms like "India," "summit," "Modi," "Russia," and "Ukraine." In contrast, IndianExpress delves into discussions possibly about India's stance in the global health or technological sectors, highlighted by terms such as "India," "health," "global," "digital," and "challenge." Lastly, the NYTimes seems to

spotlight interpersonal dynamics and relations between global leaders, with a specific focus on discussions between "Biden" and "Modi," underpinned by terms like "Biden," "country," "Modi," "leader," and "India."

Comparing text statistics



Metric	BBC	NYTimes	IndianExpress
Word Count	3041.00	5015.00	1656.00
Line Count	186.00	236.00	61.00
Average Word Length	4.42	4.55	4.69
Average Line Length	85.72	114.95	150.10
Vocabulary Size	971.00	1486.00	657.00
Lexical Diversity	0.32	0.30	0.40
Longest Word	ground-breaking...	democracy-versus-autocracy...	Chinese-chequered...
Longest Line	"It was clear...	The project lacked...	Beyond an obvious need...
Average Sentence Length	142.29	187.22	203.22
Number of Sentences	112.00	145.00	45.00

Code overview
The script meticulously analyzes textual statistics for various news sources. By tokenizing the texts into words and sentences and evaluating them in distinct ways, it captures attributes such as word count, line count, and lexical diversity. Once the texts from different news sources are combined and processed, these statistics provide insight into their respective linguistic structures and patterns.

Analysing the statistics

BBC's content, with 3,041 words and a lexical diversity of 0.32, suggests a balanced blend of repetitiveness and vocabulary variation. In contrast, NYTimes offers a heftier word count of 5,015 but with a slightly lower lexical diversity of 0.30, indicating a denser but more repetitive narrative. IndianExpress, despite its smaller word count of 1,656, boasts the highest lexical diversity at 0.40, hinting at a richer vocabulary usage in a concise format. The longer sentences in IndianExpress, averaging 203.22 characters, reflect a more in-depth elaboration per statement, while BBC and NYTimes present more succinct, direct points. The differing vocabulary sizes and sentence lengths across these outlets mirror their unique editorial nuances and content presentation styles.

THANK YOU PROFESSOR BIPLAV!