Introduction to Text Mining and Natural Language Processing

Homework-2

Submitted by:

Deepak Kumar Malik,
Iñigo Exposito,
Enzo Infantes

Barcelona School of Economics
Master in Data Science for Decision Making
Master in Data Science for Methodology

February 25, 2025

Contents

1 Part-1 Getting Main Text Data

Our database was sourced from Kaggle (link to our dataset) and it consists of news articles collected over the past few months using the NewsAPI, aimed at supporting the development and testing of NLP models for text summarization, sentiment analysis and other applications. Sourced from a wide range of reputable outlets, the dataset includes article texts covering a wide variety of topics (making it suitable for our text analysis), publication dates, source information, and additional metadata, covering a wide variety of topics. The dataset contains the following columns, with descriptions provided where available:

Column Name	Description
`article_id`	Unique article ID
`source_id`	Source title
`source_name`	Source name
`author`	The author of the article
`title`	The headline or title of the article
`description`	A description or snippet from the article
`url`	The direct URL to the article
`url_to_image`	The URL to a relevant image for the article
`published_at`	The date and time the article was published (in UTC)
`content`	The unformatted content of the article, truncated to 200 characters
`category`	The type of article (politics, economics, health)
`full_content`	The unformatted content of the article

Some columns, such as `source_id`, `url`, `url_to_image`, and `content`, were excluded from our analysis as they were not relevant for our purposes. `Source_id` was redundant since we already have `source_name`, and `url` and `url_to_image` were not used in our analysis. Instead of using the `content` column, we opted to work with the `full_content` column, which contains the complete article text rather than just a snippet. After removing those columns, we dropped duplicates in order to avoid having the same entry more than once. Once we performed that task, we dediced to handle NA values for different categories.

full_content	author	description	category	title
0.460985	0.080711	0.003732	0.000412	0.000393

The previous table summarizes the proportion of missing values among different categories. For `author`, `description`, `category`, and `title`, we decided to recategorize missing values by creating the following categories, respectively: `unknown`, `uncategorized`, `no title`, and `no description`. The highest number of missing values was found in the `full_content` column. To maintain data quality, we decided to remove those entries with missing content, as they would not be useful for analysis. Out of 101,832 total articles, 54,889 contain valid `full_content`, whereas 46,943 articles have missing `full_content`. As a result, our analysis is based on a total of 58,432 articles, as seen in the following plots.

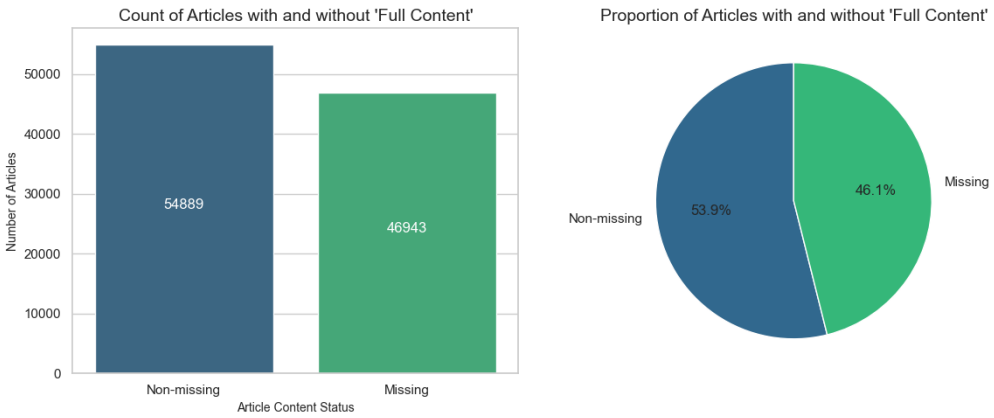

Figure 1: Distribution of `full_content` missing values

Text Processing Pipeline

Once we have cleaned and understood our dataset, we will proceed with some preprocessing by following these steps. All this tasks were implemented within the `process_text_pipeline` function.

- **Lowercasing**: Converts text to lowercase for consistency.
- **Date Removal**: Uses regex to remove various date formats.
- **Special Character Removal**: Removes punctuation and non-alphanumeric characters.
- **Tokenization & Lemmatization**: Uses spaCy to tokenize and lemmatize text.
- **Stopword Removal**: Removes common and news-related stopwords.
- **Parallel Processing**: Uses swifter.apply() for speed optimization.

Once we have understood the structure of our text, we will plot the number of articles by source name, the number of articles over time and the number of articles amongst each category. Let us begin by plotting the number of articles produced by each newspaper, which could help detect biases, analyze trends, compare coverage and evaluate data reliability.

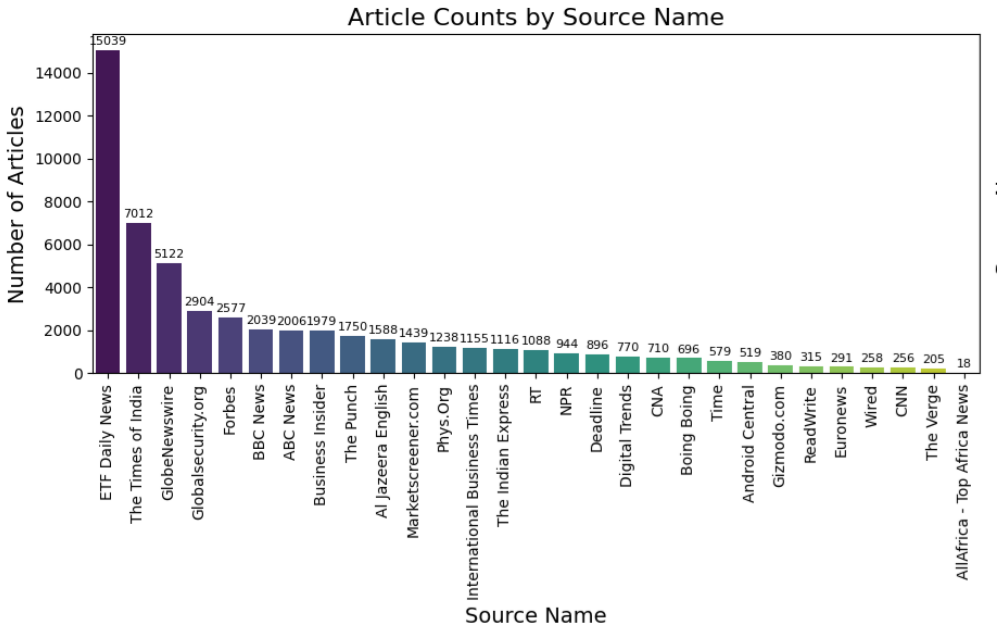

Figure 2: Article counts by source name

The previous bar chart displays the number of articles per source, showing a highly uneven distribution. ETF Daily News dominates with 15,039 articles, followed by The Markets Insider (7,012) and Cointelegraph (5,122). A steep decline follows, with most sources contributing under 3,000 articles. The lower end of the spectrum includes sources like The Verge and AllAfrica - Top Africa News, each with fewer than 300 articles.

By plotting the number of articles over time (daily or weekly), we can analyze the evolution of trends and identify whether certain dates hold greater significance due to important events. This approach enables a more comprehensive and informed analysis.

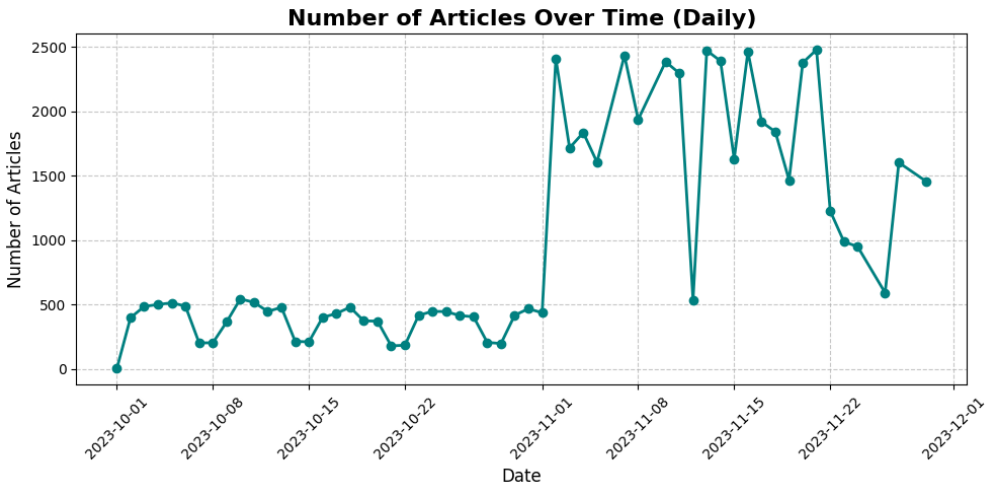

Figure 3: Article counts by days

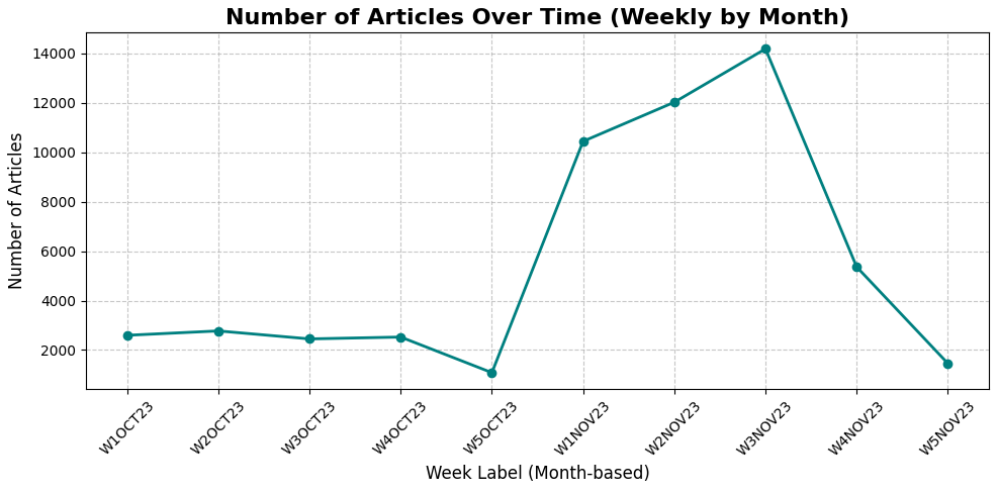

Figure 4: Article counts by week

It is challenging to extract meaningful insights from the previous plots. However, one clear observation is a significant increase in publications around November 11. Despite this, no useful trends are immediately apparent. We will conduct a deeper analysis by examining how each topic evolves over this time period. Finally, by summarizing the data by category, we can gain insights into which categories have the most influence on the articles. This will allow us to identify dominant themes and trends, highlighting areas of significant focus throughout the period.Due to the high number of categories, we decided to focus on plotting only those categories with more articles than the average. This approach allows us to highlight the most prominent categories and better understand the areas with the greatest concentration of publications.

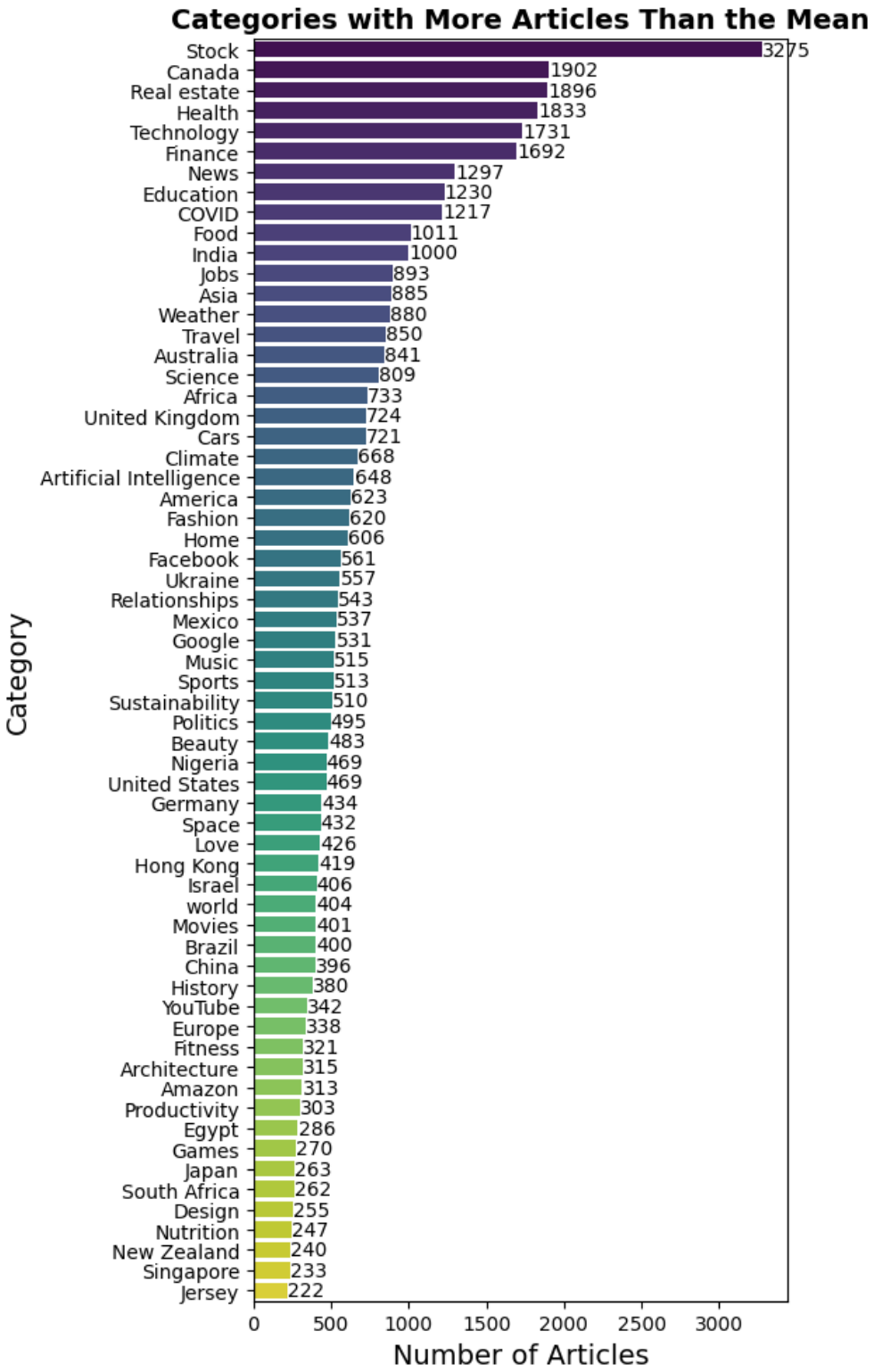

Figure 5: Categories with more articles than the mean

We will conclude this subsection by presenting two word clouds. The first one represents categories with more articles than the average, while the second covers all articles. The differ-

ences between the two word clouds are not very noticeable, indicating a similar distribution of key terms across both groups.

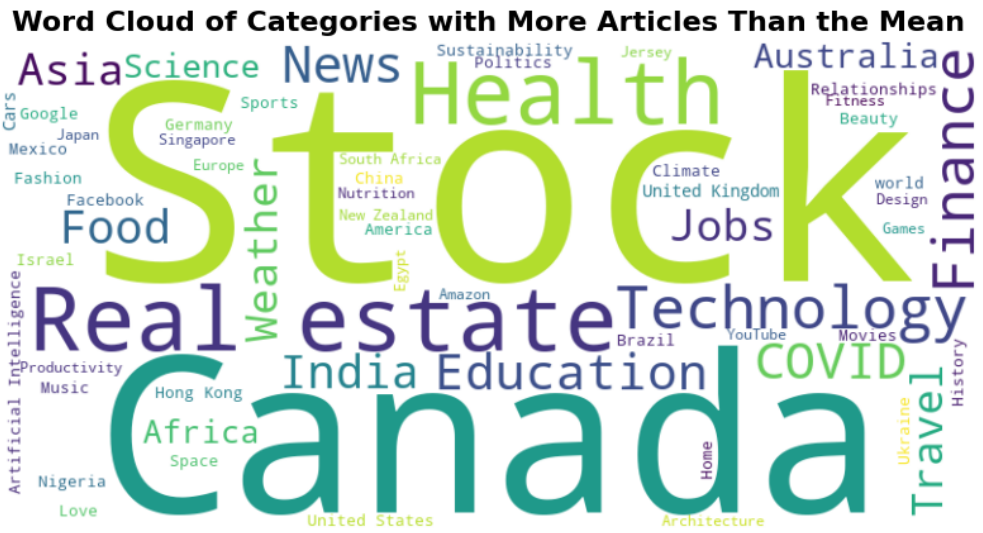

Figure 6: World cloud of categories with more articles than the mean

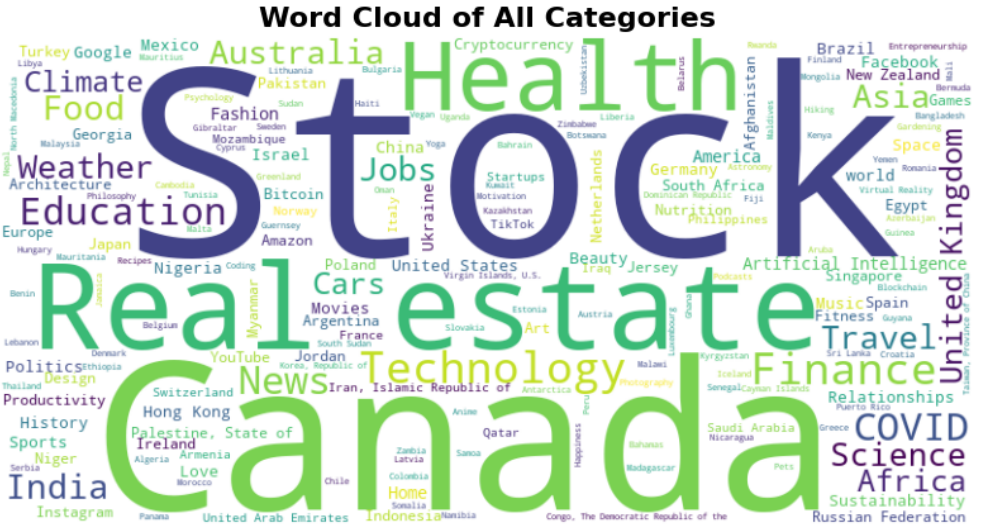

Figure 7: World cloud of all categories

As mentioned earlier, the large number of categories complicates our analysis. To address this, we will perform a recategorization by creating a new category column. The following pipeline will be followed:

1. **Category Mapping**: Define broad categories and map specific keywords to these categories.

2. **Categorizing Articles**: Each article will be assigned to one of the broad categories based on its current category. Articles that do not match any predefined category will be labeled as `Uncategorized`.

3. **Cleaning**: A new column, `Broad_category`, will be added to the dataset, and rows labeled as "Uncategorized" will be removed.

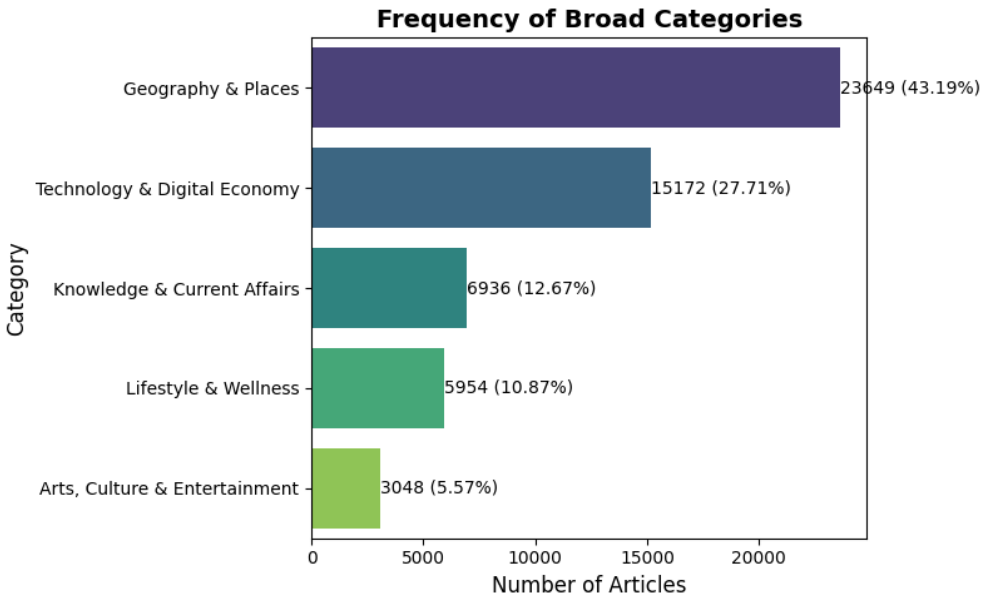

Figure 8: Distribution of Broad Categories

With this recategorization, we aim to identify patterns around categories, dates, and the number of articles. Let us begin by analyzing the percentage distribution of articles by category. The following plot shows that by the end of October, the "Geography and Places" category gains significant importance. This is likely due to the events surrounding the October 7th attack by Hamas in Israel, which resulted in over 1,400 Israeli casualties, the majority of whom were Jewish, alongside several hundred others of different nationalities. The violence also caused numerous injuries and had a devastating impact on both Israeli civilians and Hamas fighters. Given this, we expect words like "Israel," "Palestine," and "Hamas" to be prominent.

In November 2023, several key events shaped the financial landscape. The first anniversary of ChatGPT marked significant advancements in AI, influencing sectors like education and business. On Wall Street, investor confidence remained high, particularly in the technology sector, driven by strong earnings and optimism around innovation. A more accommodative Federal Reserve policy, with lower interest rates and a dovish stance, supported economic growth and boosted investor sentiment, especially in tech stocks. These factors combined for a standout month in the markets, and we anticipate words like "stock," "Wall Street," and "NYSE" to appear frequently.

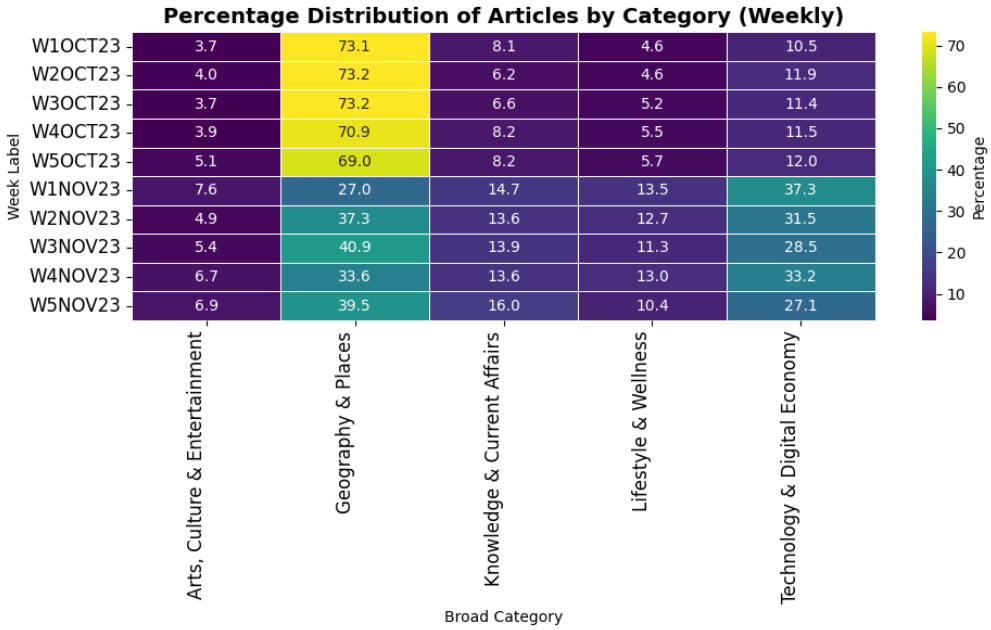

Figure 9: Weekly distribution of articles by broader categories

We can confirm that our intuition holds true by analyzing the results of the following word clouds for the topics mentioned above.

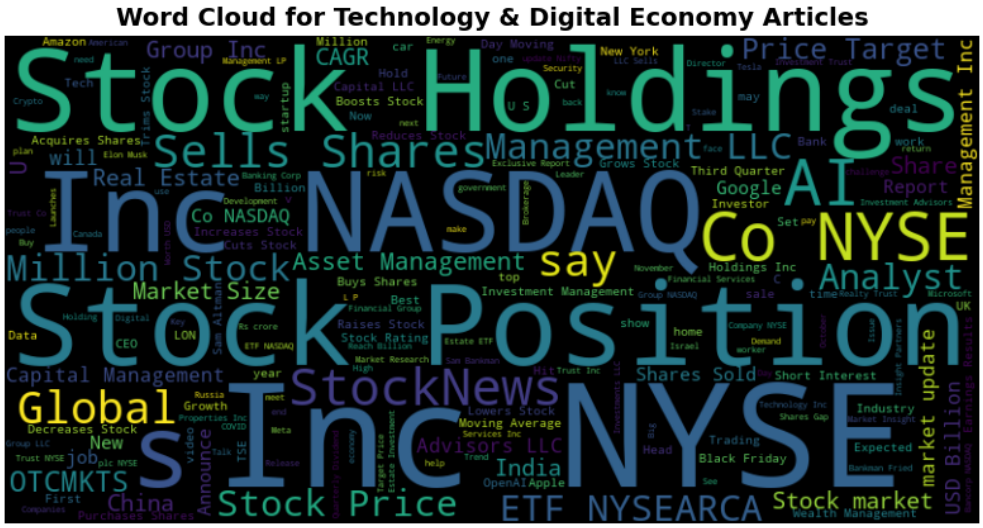

Figure 10: Technology and Digital Economy

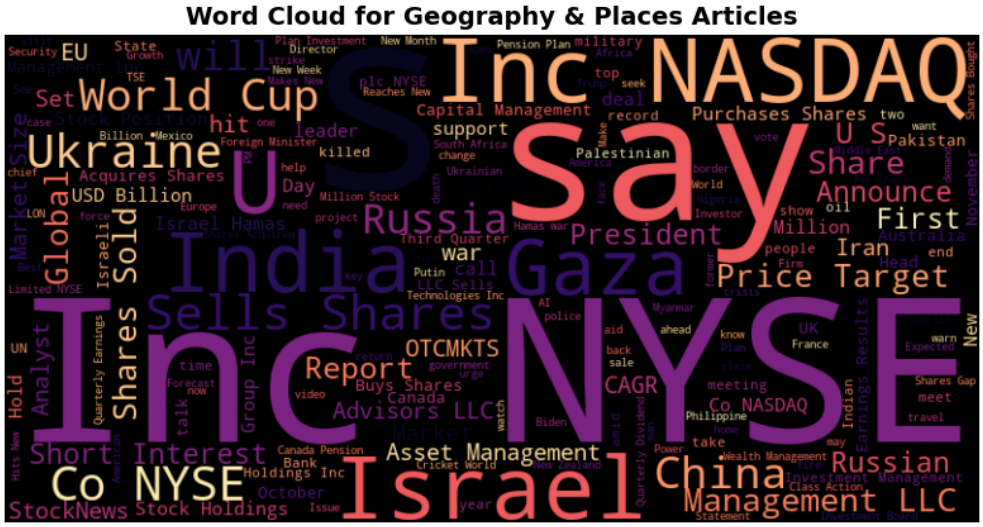

Figure 11: Geography and Places

The final preprocessed dataset (`final_preprocessed_data`) contains the following columns:

- **article_id**
  - **Description:** Unique identifier for each article.
  - **Usage:** Useful for tracking and linking articles across analyses.
- **source_name**
  - **Description:** The name of the news source that published the article.
  - **Usage:** Allows grouping and analysis by news outlet to study media bias or source-specific trends.
- **published_at**
  - **Description:** The publication date and time of the article (converted to a date-time format).
  - **Usage:** Facilitates time-based analysis, such as trends over days, weeks, or months.
- **full_content**
  - **Description:** The original text content of the article, if available.
  - **Usage:** Serves as the raw text from which the processed content is derived.
- **processed_content**
  - **Description:** The cleaned and preprocessed version of the article text. This includes steps like lowercasing, punctuation removal, tokenization, stopword removal, and stemming/lemmatization.

  - **Usage:** Used for subsequent text analysis tasks such as dictionary generation, topic modeling, and sentiment analysis.

- **processed_preview**
  - **Description:** A short preview (first 200 characters) of the `processed_content`, with an ellipsis appended if the text is longer.
  - **Usage:** Provides a quick visual check of the preprocessed text without displaying the full content.

- **category**
  - **Description:** The topic or category assigned to the article (e.g., Politics, Technology, Health).
  - **Usage:** Enables grouping and comparative analysis of articles by topic.

- **week_start**
  - **Description:** The starting date (typically Monday) of the week in which the article was published.
  - **Usage:** Useful for aggregating articles on a weekly basis for time-series analysis.

- **week_label**
  - **Description:** A custom label representing the week within a month (e.g., "W1OCT23"), created using the publication date.
  - **Usage:** Provides an intuitive and human-friendly way to view and compare weekly article trends.

- **week_order**
  - **Description:** A numeric value combining the year, month, and week number (calculated as `year*10000 + month*100 + week_in_month`) used solely for sorting weekly summaries in chronological order.
  - **Usage:** Ensures that weekly data is correctly ordered when visualizing trends over time.

According to the requirements in homework 02, we will show some examples of our text before and after the preprocessing steps. In the following table we can see that difference:

Comparison of Original and Processed Sentences	
Original Sentence	**Processed Content**
UN Secretary-General Antonio Guterres urged the world Monday to "stop the madness" of climate change as he visited Himalayan regions struggling from rapidly melting glaciers to witness the devastating impact of the phenomenon.	secretary general antonio guterres urge world monday stop madness climate change visit himalayan region struggle rapidly melt glacier witness devastating impact phenomenon
Cofounder at UpperKey. Passionate about property management, real estate investments, proptech and driving international business growth. There are a number of reasons investors might consider real estate investments in Dubai.	cofounder upperkey passionate property management real estate investment proptech drive international business growth number reason investor might consider real estate investment dubai
India, the first non-Arab country to recognise the PLO in the 1970s, is now seen closer to Israel and its biggest benefactor, the United States. New Delhi, India–Israel's relentless bombing of the besieged Gaza Strip and killing of nearly 6,000 people – a third of them children – in two weeks.	india first non arab country recognise plo 1970s see closer israel big benefactor united state delhi india israel srelentless bombingof besiege gaza strip kill nearly people third child two wee

In the previous table, the left column displays brief original sentences from our data. The right column contains the same sentences after undergoing our preprocessing steps. In the preceding section, we discussed the considerations for these steps (using the *process_text_pipeline* function).

2 Develop and implement methodology

After the deep exploratory analysis conducted in the previous section, we developed the following research question: *"How did media narratives evolve from a predominantly geopolitical focus in October 2023 to a more technology-driven discourse in November 2023, and what specific linguistic markers distinguish these shifts in coverage?"*

Our analysis of the Global News Dataset for October–November 2023 revealed a striking temporal transition in the topics covered by media outlets. In October, our dictionary-based analysis indicated that "Geography & Places" dominated the news content, with approximately 70%–73% of articles focusing on geopolitical issues. However, by November, there was a notable shift in the distribution, with "Technology & Digital Economy" emerging strongly, accounting for 30%–37% of the content. This shift coincided with a decline in the emphasis on geographic and geopolitical reporting.

This transition aligns with real-world events during the period:

- The escalation of the Israel-Hamas conflict in early October likely contributed to the heavy focus on geographical and geopolitical reporting.
- In contrast, early November witnessed significant technological milestones, such as the release of GPT-4 Turbo by OpenAI and Meta's launch of Threads, a competitor to Twitter.

Our research question thus aims to uncover not only the extent of this narrative shift but also to identify the linguistic markers—via dictionary-based methods—that characterize the transition from geopolitical to technology-centric reporting. This investigation promises to provide valuable insights into how external events influence media framing and the evolution of public discourse. We will follow two different methodologies.

Methodology 1: Garcia-Uribe et al. (2023)

In alignment with our research question, we will focus on the content of the articles to identify the most frequent words within the categories of "Geography & Places" and "Technology & Digital Economy." To achieve this, we will adjust the vectorizer specifically for geopolitical and technology-related content, given that our analysis is focused on only these two types of text. We will then present the top 20 most frequent words in each category. The following are the top terms identified in the "Geography & Places" and "Technology & Digital Economy" categories:

Term	Frequency
buy	1228.32
ratio	1174.32
average	1100.59
israel	1075.79
india	1042.08
people	999.27
country	976.97
value	948.55
dividend	927.06
world	900.32
service	831.76
sell	829.99
additional	809.31
target	799.44
llc	783.71
worth	741.76
government	733.34
security	722.82
say	717.89
president	716.40

Table 4: Geography & Places

Term	Frequency
dividend	910.77
llc	894.91
transaction	693.18
bank	670.16
earnings	626.29
stake	622.34
october	607.64
purchase	605.22
holding	595.10
india	569.78
use	565.97
trade	548.22
acquire	522.11
real	515.51
product	514.67
million	514.51
capital	505.75
trust	503.22
global	495.91
november	489.61

Table 5: Technology & Digital Economy

The objective of this project is to analyze the evolution of media narratives, transitioning from a geopolitical focus in October 2023 to a more technology-driven discourse in November 2023. A TF-IDF analysis of the Global News Dataset for these two months reveals a significant shift in thematic emphasis.

- Geography & Places

  Top Terms: buy, ratio, average, Israel, India, people, country, value, dividend.
  Interpretation: The frequency of terms such as `Israel`, `India`, `country`, and `people` reflects a predominant focus on geopolitical events and geographic locations. The inclusion of terms like `buy`, `ratio`, `average`, and `value` suggests that economic considerations are being integrated into the discourse surrounding geopolitical contexts. Additional terms such as `government`, `security`, and `president` further underscore the political and governance aspects prevalent in the media coverage during this period.

- Technology & Digital Economy

  Top Terms: dividend, LLC, transaction, bank, earnings, stake.
  Interpretation: The prominence of terms like `LLC`, `transaction`, `bank`, `earnings`, and `stake` signifies a notable shift towards discussions centered on financial investments, business structures, and economic transactions. Terms such as `purchase`, `holding`, and `trade` indicate a heightened focus on market activities and capital

flows within the technology sector, further corroborating the shift from geopolitical to technology-driven content.

Methodology 2: Hassan et al (2019)

In line with our research question, we focus on the content of the articles to identify the most frequent terms in the categories of Geography & Places and Technology & Digital Economy. In this analysis, we will adjust the vectorizer specifically for geopolitical and technology-related content. To align with the methodology proposed by Hassan et al. (2019), we will employ bi-grams as the basis for the frequency analysis. This approach allows for the capture of more meaningful and contextually relevant terms compared to unigrams (single words), especially in the context of complex topics like geopolitics and technology. The results will include the top 20 bi-grams for each category, providing a clearer understanding of the linguistic patterns associated with each thematic shift.

Bigram	Frequency
institutional investor	1525.78
hedge fund	1360.07
day average	1345.21
united state	1325.35
purchase additional	1163.68
buy hold	1126.21
additional period	1027.20
best buy	1024.62
want great	1017.19
idea sell	1015.09
sell double	1015.09
yearthe best	1014.93
double yearthe	1014.93
invest idea	1014.93
great invest	1014.93
earnings share	986.82
average day	952.74
acquire additional	889.43
security exchange	850.35
ratio beta	845.18

Table 6: Geography & Places

Bigram	Frequency
real estate	1668.38
institutional investor	1594.14
hedge fund	1440.90
day average	1382.45
purchase additional	1223.42
buy hold	1102.66
additional period	1082.47
best buy	1080.54
sell double	1026.66
idea sell	1022.09
want great	1021.92
invest idea	1021.74
yearthe best	1021.57
great invest	1021.57
double yearthe	1021.57
average day	988.39
acquire additional	972.77
security exchange	900.98
buy additional	899.75
earnings share	895.54

Table 7: Technology & Digital Economy

Our project aims to track the evolution of media narratives from a geopolitical focus in October 2023 to a technology-driven discourse in November 2023. The CountVectorizer analysis using the Hassan et al. (2019) methodology of the Global News Dataset for these months reveals a noticeable shift in focus:

- Geography & Places

  Top bigrams: institutional investor, hedge fund, day average, united state, purchase additional.
  Interpretation: The prominence of bigrams like "institutional investor," "hedge fund," and "day average" reflects a strong emphasis on economic and financial contexts within geopolitical reporting. The presence of "united state" emphasizes geopolitical locations, and terms like "purchase additional" suggest discussions around transactions and investments.

- Technology & Digital Economy

  Top bigrams: real estate, institutional investor, hedge fund, day average, purchase additional.
  Interpretation: The presence of bigrams like "real estate," "institutional investor," and "hedge fund" indicates a focus on financial and investment aspects within the technology sector. The bigram "day average" points to discussions on market analysis, and "purchase additional" suggests a focus on transactions and acquisitions.

In October, the focus was predominantly on geopolitical situations, with significant coverage of financial and investment aspects related to geopolitical events. By November, the narrative shifted towards technology and digital economy, emphasizing real estate, financial investments, and market analysis within the technology sector.

To achieve these results, we needed to adjust some parameters within our functions. For example, in the first methodology, we used TfidfVectorizer() to create our dictionary with the most frequent tokens according to the Tf-idf approach. In this case, we used English as our stop words language, set the maximum percentage that a word can appear in the entire corpus to 0.4, and the minimum to 0.1. This is because we are trying to remove common words in niche datasets for balanced coverage. The same idea applies to the second methodology when using CountVectorizer().

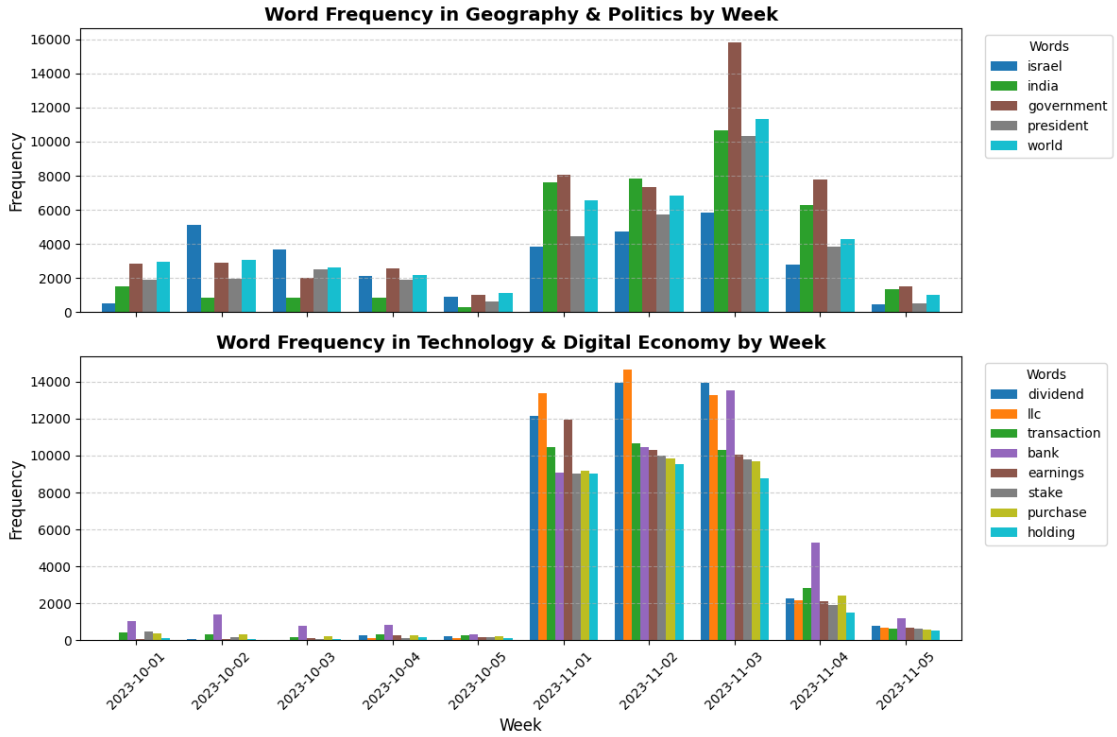

Figure 12: Top words with more sense in each category by week

The bar chart visualization offers a sharp view of the frequency of key terms in geopolitical and economic contexts, with each chart providing insights into different aspects of the global landscape. In the top chart, terms such as "Israel," "India," and "government" highlight the intense political focus surrounding the Hamas-Israel conflict in early November 2023, marking a notable surge in discussions around governance and international relations. The continued presence of these terms after the peak suggests that while media attention may have decreased, political discussions continued to shape public discourse. This temporal shift highlights how certain geopolitical events can sustain a level of discourse well beyond their immediate crisis period. The bottom chart, focusing on the technology and digital economy, illustrates the fluctuations in financial terminology such as "dividend," "LLC," and "transaction." The spike in early November 2023 correlates with the U.S. stock market rally, demonstrating how financial markets, corporate earnings, and investor sentiment are key drivers in shaping media narratives. The prominent mentions of terms like "bank" and "earnings" reflect ongoing discussions about economic activity and market movements.