Report Project UDA

CID: 02354789

In this detailed study, we set out to decode and contrast the global news cycle across specific one-week intervals in 2017 and 2018. Specifically, from 24th August to 30th August. The core aim is to deconstruct the dynamics of worldwide news coverage, pinpointing shifts in themes and detecting variations in sentiment and focus during these times. Utilizing a dataset named "Global News Week" from Kaggle [1], which aggregates headlines from multiple sources over one week in each of the two years, we plan to shed light on the patterns and tendencies defining the media narrative in these periods. The dataset contains a vast array of over one million records, including publication dates and headline texts, offering an extensive snapshot of the media environment. However, due to computational constraints and the need to keep the dataset size manageable at approximately 100MB, our analysis will concentrate on a subset of records from each year. This selection allows us to maintain a broad and representative sample while ensuring the feasibility and efficiency of our computational analysis.

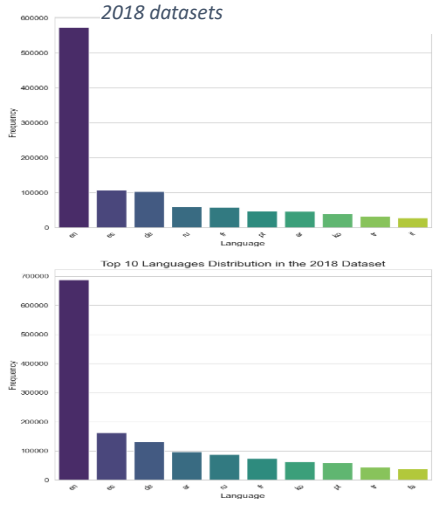
Problem Statement: *Our objective is to conduct a thorough exploratory data analysis coupled with advanced text analytics like Topic Modeling, Topic Clustering and Sentiment Analysis to provide a comprehensive depiction of the impact of major global events and dominant narratives on the temporal and thematic structure of news coverage. Specifically, we seek to compare the influence of these elements during the weeks of August 24-30 across the years 2017 and 2018, highlighting any shifts or continuities in media focus and presentation.*

I. Data Selection and preparation

i. Understanding and Selecting the Data

This dataset comprises a compilation of global news articles from a specific week in each 2017 and 2018, detailing the publication time, feed code, source URL, and headline text. Each record documents a distinct news event, ranging from local incidents, such as tree falls, to broader topics like international alerts and business achievements. Organized in a table, the data offers a glimpse into the various subjects and happenings reported by diverse news platforms globally, showcasing the range and depth of worldwide news coverage during these periods. The headlines are presented in multiple languages. Using the *langdetect* library [3], we analyzed and identified the language for each headline. Figure 1 indicates that a predominant number of headlines in both datasets are in English. For efficiency in processing, we have chosen to focus solely on English headlines for subsequent analyses. This decision notably reduces the size of the datasets, ensuring their combined compressed size remains under 100mb.

Figure 1 Top 10 languages in 2017 and 2018 datasets



ii. Data Cleaning and Preprocessing

In the initial phase of data preparation and preprocessing for the global news articles dataset from 2017 and 2018, the 'publish_time' column undergoes transformation from a string to a datetime format, allowing for the detailed breakdown of temporal components such as year, month, day, hour, and minute. This step is crucial for enabling time-series analysis and understanding trends over time. Following this, the dataset is further refined using Natural Language Processing techniques. Essential Natural Language Toolkit (NLTK) resources are employed, including stopwords for eliminating common words that offer little value, WordNet Lemmatizer for condensing words to their root form, and a tokenizer for breaking down text into individual terms. The cleaning function meticulously processes each headline, standardizing the text to lowercase, stripping non-alphabetic characters, and applying both tokenization and lemmatization while removing stopwords. This rigorous cleansing ensures that the headlines are devoid of extraneous information and are uniform in structure, making them suitable for in-depth linguistic and textual analysis. The result is a set of clean, normalized headlines, optimized for subsequent analytical tasks and insights extraction.

II. Exploratory Data Analysis (EDA)

i. Descriptive and Statistical Analysis

In our descriptive and statistical analysis of the 'cleaned_headline' texts from 2017 and 2018, several key insights emerged. The dataset for 2018 contained a larger count of articles at 688,563 compared to 573,721 in 2017, indicating a higher volume of reported news in English in 2018. The mean values for both years suggest that headlines in 2018 were slightly longer, with a mean of 53.10 compared to 51.76 in 2017. Despite this, the median values for both years were close to the mean, suggesting a symmetric distribution around the central tendency. The range was notably larger in 2018, indicating greater variability in headline length or composition. However, the variance and standard deviation were slightly higher in 2017, pointing to a more pronounced spread in the data for that year. These statistics collectively provide a nuanced understanding of the headline text's characteristics over the two years, highlighting subtle shifts in news reporting volume and headline complexity, which could reflect broader trends in news coverage and reporting practices during this period.

ii. Temporal Analysis

Since our datasets also provide publication time, it is also interesting to compare the two datasets in terms of volumes of publication daily during weeks days and weekends as well as hourly.

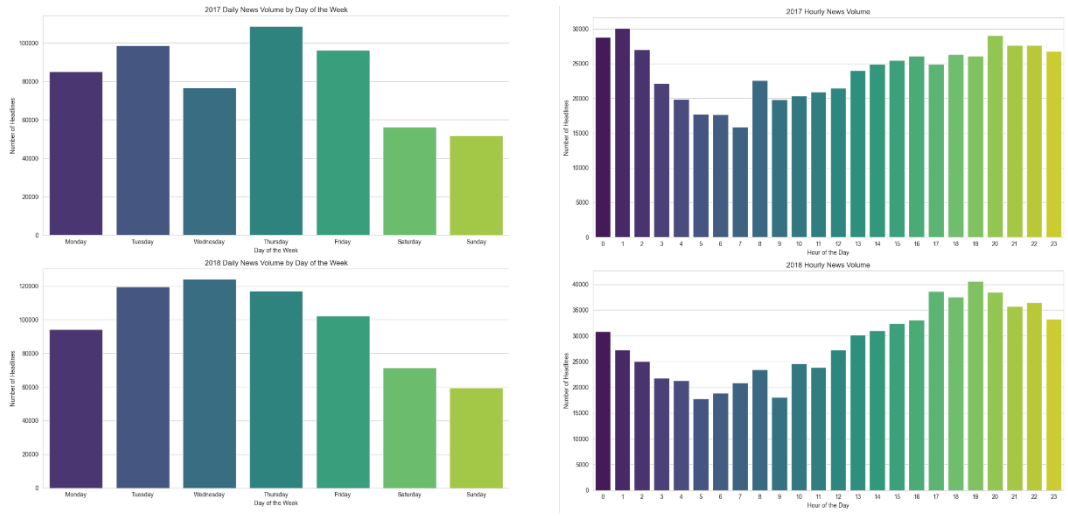


Figure 2 Daily and Hourly publication volume in Week of August 2017 and 2018

As depicted in Figure 2, it becomes apparent that the daily volume distribution closely resembles each other during the week of August in both 2017 and 2018. However, a noteworthy exception arises on Wednesday, where a substantial surge in news volume occurred in 2018. This anomaly suggests the possibility of a significant event taking place on that particular day. Turning our attention to the hourly volume distribution, it's evident that the patterns remain largely consistent across both years with an increase in publication in the evening and early night.

iii. Frequency analysis

The examination of 'cleaned_headline' texts from 2017 and 2018 through frequency analysis has offered valuable insights into the key themes and subjects prevalent in worldwide news outlets over these years. The technique employed to ascertain the frequency of words involves tokenizing the aggregated text of headlines and then counting the occurrences of each word. In 2017, the frequent appearance of terms such as "market," "Trump," and "Harvey" signals a concentration on financial matters, the political narrative around the president at the time, and notable incidents like Hurricane Harvey. The 2017 word cloud (Figure 3) highlights these themes with more pronounced and recurrent depictions of

Figure 3 Word Clouds based on frequency in 2017 and 2018

related words. On the other hand, 2018's analysis not only continues to feature words like "market" and "Trump" but also shows a rise in terms like "inc" and "share," suggesting a sustained focus on economic and political news, potentially with increased reporting on corporate affairs and the stock market.

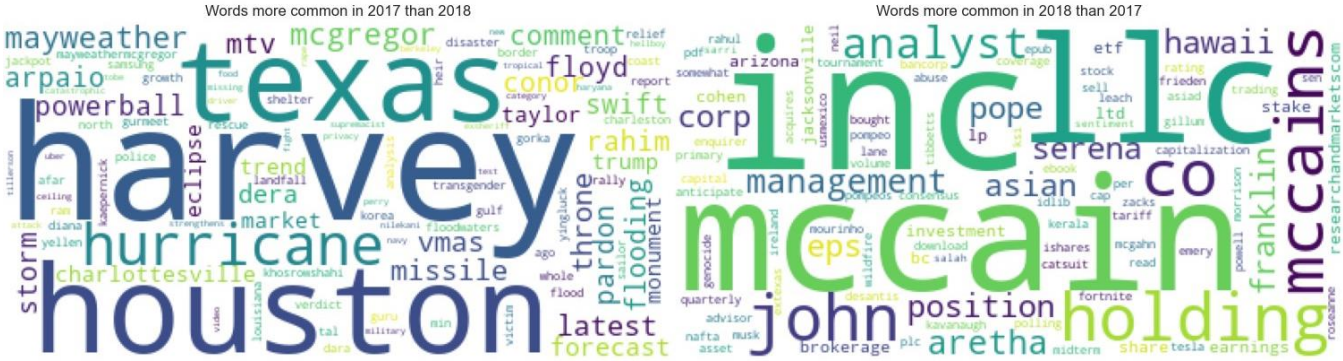


Figure 4 Word Clouds based on TF-IDF for 2017 and 2018

In this analytical approach, we harnessed TF-IDF (Term Frequency-Inverse Document Frequency) scores to uncover words that exhibited significant differences in usage between 2017 and 2018. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents, the corpus. It increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. By applying a discerning threshold of 0.000005, we meticulously curated word clouds for both years, visually accentuating these distinctions. This methodology offered us a profound insight into the evolving linguistic landscape. A comparative analysis with previous word clouds underscored the specificity of headlines in each dataset. Notably, we observed that while words like "Trump" and "market" were prevalent in both years, TF-IDF analysis brought to the fore the nuanced shifts. For instance, the prominence of "Hurricane Harvey" in August 2017 and "John McCain" in August 2018 exemplifies the temporal focus of news coverage during those periods. This comprehensive approach enhances our understanding of textual dynamics across the two years.

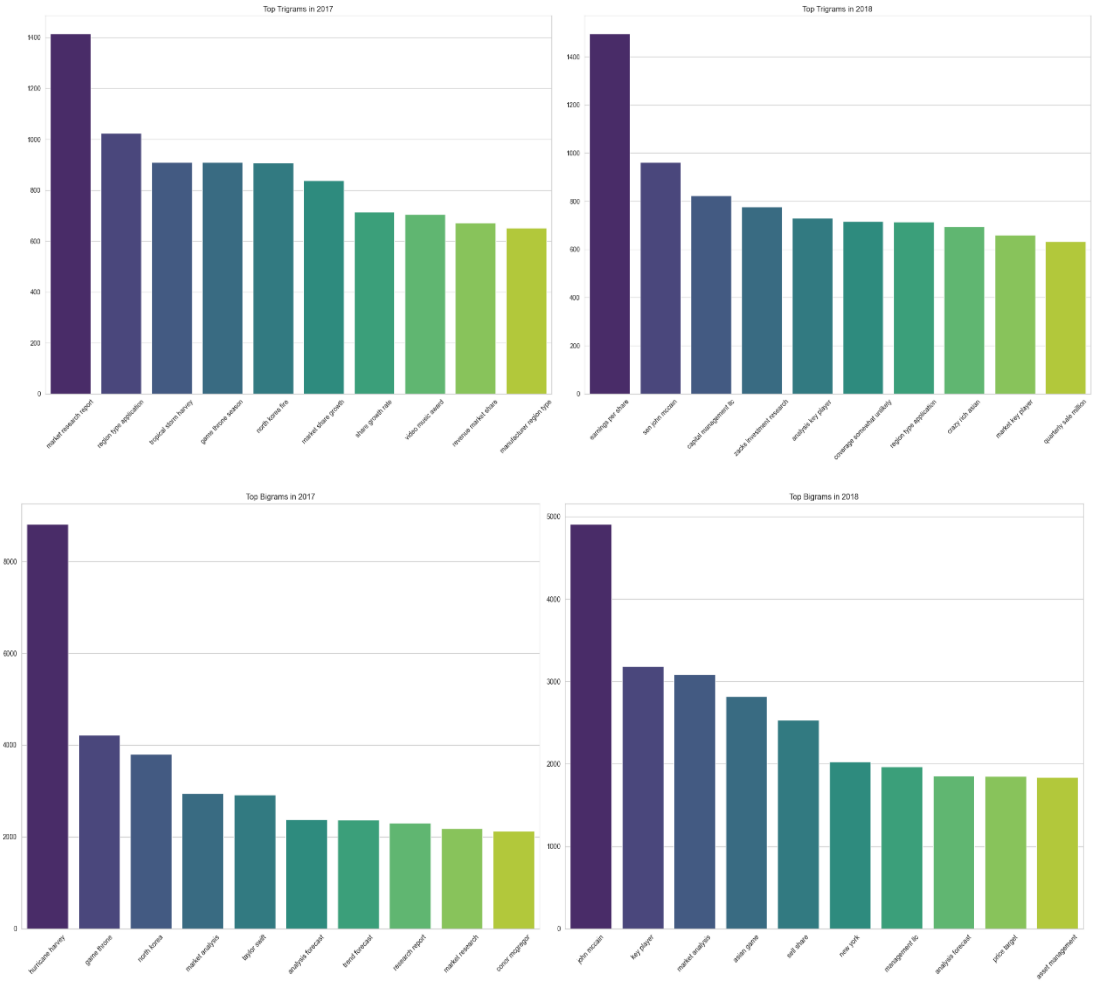


Figure 5 Most Common Bigrams and Trigrams in 2017 and 2018

The in-depth analysis of bigrams and trigrams, which are types of n-grams representing sequences of two and three contiguous words, offered a more nuanced understanding of the prevailing narratives and focal points in each year's media attention. This n-gram analysis, extracting these word sequences through a sliding window technique over the text, does not adjust for the commonality of terms across the corpus, allowing unique, context-specific terms, often proper nouns and specialized vocabulary, to emerge. In 2017, bigrams such as "hurricane Harvey" and "Taylor Swift" pointed to major events and notable personalities, while trigrams like "market research report" and "game throne season" suggested specific interests in business and entertainment. This indicates a year dominated by discussions of both disaster and popular culture. In contrast, 2018's bigrams and trigrams, with terms like "John McCain," "Asian game," "market analysis," and "asset management," reveal a shift towards diverse topics spanning politics, international events, and financial news. The consistent presence of geopolitical themes, as seen in trigrams such as "ser John McCain," highlights the ongoing political dialogues. These detailed linguistic insights not only spotlight the distinct individuals, events, and themes that were news highlights in each year but also depict the evolving landscape of public interest and media focus, illustrating the dynamic and shifting nature of news narratives across political, economic, and social domains.

iv. Temporal/Frequency Analysis

Figure 6 Word Clouds per day in the week of 2017 and 2018

The word clouds displayed in the image reflect a temporal analysis of two global news datasets, each representing a distinct week in August for the years 2017 and 2018. The visualization captures the prominence of specific topics on each day, revealing the fluctuating landscape of news coverage. In 2017, the beginning of the week spotlights financial markets, as seen by the word "market" across the first few clouds. Midweek, the emergence of "Harvey" suggests a shift in news focus, likely due to Hurricane Harvey's impact. By the end of the week, "Harvey" becomes even more dominant, indicating ongoing coverage of the disaster's developments. For 2018, "market" and "new" remain consistently significant, implying a steady stream of financial news throughout the week. However, the appearance of "McCain" midweek indicates a spike in news regarding Senator John McCain, which could be associated with a notable event involving him at that time. The variation in word sizes day-to-day underscores the ebb and flow of news topics as they gain or lose prominence. The TF-IDF methodology applied enhances the representation of the most pertinent topics for each day, rather than those most frequently mentioned. This nuanced temporal analysis underscores the day-specific attention that different news topics receive, providing a clear snapshot of the week's news cycle dynamics.

v. EDA Conclusion

The exploratory data analysis of 'cleaned_headline' texts from 2017 and 2018 has unveiled a dynamic news landscape, marked by an increase in article volume and headline length in 2018, suggesting a richer and more detailed reporting environment. The temporal and statistical insights reveal subtle shifts in news presentation and focus, with 2017 headlines showing more dispersion, and 2018 characterized by greater variability in length. The frequency analysis, incorporating bigrams, trigrams, and TF-IDF scores, highlights the evolving thematic interests, from "Hurricane Harvey" and "Taylor Swift" in 2017 to "John McCain" and "Asian games" in 2018. This comprehensive EDA underscores the nuanced nature of global news narratives, reflecting shifts in media practices, public interest, and the world's socio-political climate, providing a foundation for understanding and anticipating future trends in news reporting.

III. Advanced Text Analysis

Exploratory Data Analysis (EDA) and basic textual examination have provided preliminary insights into the main subjects and news themes for each week in the years 2017 and 2018. To deepen our understanding and draw a thorough comparison, we will engage in Advanced Text Analysis by implementing Topic Modeling and Sentiment Analysis, which will aid in identifying the core narratives and assessing the sentiment embedded in the coverage of these global news events.

i. Topic Modeling

In my analysis of global news headlines from the weeks of 24th to 30th August in 2017 and 2018, I employed two topic modeling techniques: Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). NMF is a linear-algebraic model that factors high-dimensional vectors into a non-negative matrix and a non-negative coefficient matrix, effectively clustering the input data into distinct topics. It's particularly useful when the data is non-negative, as it often leads to more interpretable and less overlapping topics. This method proved more advantageous for my datasets as it provided clearer, more distinct topics. On the other hand, LDA is a generative statistical model that assumes each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is based on probabilistic graphical modeling which provides a more abstract and less direct method of identifying topics. However, for my datasets, LDA's topics were less distinct and harder to interpret compared to NMF. The distinct and interpretable results from NMF were preferable, offering more meaningful thematic structures that resonated better with the content and context of the global news headlines I was examining. I chose to model 20 topics for my datasets as we can observe in Tables 1 and 2, finding that this number provided a detailed yet manageable set of themes to work with.

Table 1 20 topics from NMF for the 2017 dataset

0	market forecast analysis growth trend industry share application size global
1	harvey hurricane relief category landfall gas make louisiana effort major
2	korea north missile japan launch test option south shortrange table
3	new york ceo mexico zealand time library uber business song
4	trump pardon arpaio military president donald joe ban transgender white
5	latest open people mayor official shelter powerball center county ticket
6	game throne season finale watch jon question snow dragon death
7	game throne season finale watch jon question snow dragon death
8	man arrested charged shot killed murder car city accused buckingham
9	houston flood flooding water rain harvey home catastrophic floodwaters flooded
10	taylor swift video look music mtv award vmas watch single
11	say official wont china right chief russia south india expert
12	year school comment day high court state win news open
13	report global research market forecast state united health professional survey
14	texas coast strengthens flood rain prepares category visit harvey flooding
15	police suspect officer military gear arrest limit attack shooting dutch
16	storm tropical harvey category strengthens prepares watch gulf strength downgraded
17	help victim need people medium social rescue harvey flood know
18	woman missing dead death mccain navy sailor crash search john
19	food amazon price cut deal monday gas lower close drop

0	market forecast global analysis growth trend industry size application key
1	man city charged arrested utd mourinho united murder accused death
2	mccain john sen arizona senator hero dy war mccains cancer
3	new york fitness healthcare mexico minister announces australia pm prime
4	trump google house white president donald korea session north search
5	share sell stock earnings buy bought sold corp post llc
6	game asian gold video india korea medal team throne final
7	say plan study tesla musk public iran official china leader
8	shooting jacksonville dead florida tournament mass suspect video multiple authority
9	open win serena williams lead season return round th st

10	school week football high student season preview college best district
11	deal trade mexico nafta canada talk stock usmexico reach agreement
12	day labor volume trading news weekend august national start thing
13	year time prison report ago sentenced rohingya later coming period
14	million management holding llc position investment group capital stake corp
15	price target analyst somewhat news exchange hit coverage stock rating
16	state united ohio arizona california lie islamic capitol report court
17	latest hawaii hurricane lane death storm rain tropical hit franklin
18	police woman killed officer crash car arrest death suspect shot
19	pope abuse ireland sex francis church visit scandal sexual claim

Table 2 20 Topics from NMF for 2018 dataset

The application of Non-negative Matrix Factorization (NMF) for Topic Modeling has revealed a fascinating landscape of shifting news priorities and themes (Table 1 & Table 2). In 2017, the news was heavily dominated by the devastating effects of Hurricane Harvey (Topic 1), with a significant focus on its path of destruction and the ensuing humanitarian efforts. The presence of financial market discussions (Topic 0) across both years is indicative of the perennial interest in economic health, with terms like "market," "forecast," and "growth" recurring frequently. Political narratives also took center stage, with President Trump's activities, including the pardon of Joe Arpaio and military policies (Topic 4 in 2017), being heavily scrutinized. In contrast, 2018 saw a poignant shift towards memorializing Senator John McCain (Topic 2), reflecting on his contributions and legacy, which dominated the discourse. Entertainment and culture also punctuated the news, with "Game of Thrones" (Topic 6 in 2017) capturing significant media attention, underscoring the global impact of entertainment media. However, in 2018, the Asian Games (Topic 6) and the serious coverage of the Jacksonville shooting (Topic 8) illustrate a diversification in news focus, with a balance between sports and more sobering current events. A notable emergence in the 2018 dataset was the prominence of trade talks (Topic 11), pointing to the rising importance of global trade dynamics and economic policies, reflecting a world increasingly interconnected and impacted by economic treaties like NAFTA. These results from NMF Topic Modeling not only highlight the main stories and subjects that captured media attention but also show a clear shift in focus. From the natural disaster and political controversies of 2017 to the legacies of influential figures, tech industry news, and international sports in 2018, we see how global events and societal priorities evolve over time.

ii. Clustering the topics

To further understand and organize these topics, we utilized a K-Means clustering algorithm on the combined topics of 2017 and 2018. K-Means is an unsupervised learning algorithm that partitions the data into K distinct, non-overlapping subgroups (clusters) by minimizing the within-cluster variances. It operates by initializing K centroids, then iteratively assigning each data point to the nearest cluster based on the mean of the points in the cluster, and recalculating the centroids. This process continues until the positions of the centroids stabilize.We first converted each NMF-extracted topic into a dense representation using BERT embeddings. BERT (Bidirectional Encoder Representations from Transformers) embeddings are a type of deep learning model designed to understand the nuances and context of language by considering the full context of a word by looking at the words that come before and after it. This is significantly more advanced than traditional bag-of-words models which treat every word independently. The process involved tokenizing each topic, processing it through a pre-trained BERT model, and then taking the mean of the output vectors to produce a single, dense embedding per topic.These embeddings, capturing the semantic nuances of each topic, were then fed into the K-Means algorithm, specifying the desired number of clusters as 6. The algorithm iteratively assigned each topic to one of these clusters based on the proximity of its embedding to the centroids of the clusters. After convergence, each topic was assigned to one of six clusters. The resulting clusters were then analyzed to understand their thematic content. I manually reviewed the topics within each cluster and assigned labels based on the common themes among the topics they contained. This method worked effectively in my topics dataset and grouped the topics into coherent clusters that reflected the underlying patterns and themes in the global news headlines, providing a structured and insightful overview of the data. The labels assigned were: 'Natural Disasters & Weather Events', 'Market Trends & Financial Analysis', 'Cultural Events & Education', 'Global Affairs & Social Issues', 'Crime & Legal Matters', and 'Sports & Competition'.

Cluster 0: Natural Disasters & Weather Events

- harvey hurricane relief category landfall gas make louisiana effort major (2017)
- houston flood flooding water rain harvey home catastrophic floodwaters flooded (2017)
- texas coast strengthens flood rain prepares category visit harvey flooding (2017)
- storm tropical harvey category strengthens prepares watch gulf strength downgraded (2017)
- latest hawaii hurricane lane death storm rain tropical hit franklin (2018)

Cluster 1: Market Trends & Financial Analysis

- market forecast analysis growth trend industry share application size global (2017)
- report global research market forecast state united health professional survey (2017)
- market forecast global analysis growth trend industry size application key (2018)
- share sell stock earnings buy bought sold corp post llc (2018)
- deal trade mexico nafta canada talk stock usmexico reach agreement (2018)
- million management holding llc position investment group capital stake corp (2018)
- price target analyst somewhat news exchange hit coverage stock rating (2018)

Cluster 2: Cultural Events & Education

- game throne season finale watch jon question snow dragon death (2017)
- taylor swift video look music mtv award vmas watch single (2017)
- year school comment day high court state win news open (2017)
- help victim need people medium social rescue harvey flood know (2017)
- game asian gold video india korea medal team throne final (2018)
- school week football high student season preview college best district (2018)

Cluster 3: Global Affairs & Social Issues

- korea north missile japan launch test option south shortrange table (2017)
- new york ceo mexico zealand time library uber business song (2017)
- trump pardon arpaio military president donald joe ban transgender white (2017)
- latest open people mayor official shelter powerball center county ticket (2017)
- say official wont china right chief russia south india expert (2017)
- food amazon price cut deal monday gas lower close drop (2017)
- mccain john sen arizona senator hero dy war mccains cancer (2018)
- new york fitness healthcare mexico minister announces australia pm prime (2018)
- trump google house white president donald korea session north search (2018)
- say plan study tesla musk public iran official china leader (2018)
- day labor volume trading news weekend august national start thing (2018)
- state united ohio arizona california lie islamic capitol report court (2018)
- pope abuse ireland sex francis church visit scandal sexual claim (2018)

Cluster 4: Crime & Legal Matters

- man arrested charged shot killed murder car city accused buckingham (2017)
- police suspect officer military gear arrest limit attack shooting dutch (2017)
- woman missing dead death mccain navy sailor crash search john (2017)
- man city charged arrested utd mourinho united murder accused death (2018)
- shooting jacksonville dead florida tournament mass suspect video multiple authority (2018)
- year time prison report ago sentenced rohingya later coming period (2018)
- police woman killed officer crash car arrest death suspect shot (2018)

Cluster 5: Sports & Competition

- mayweather mcgregor conor floyd fight round money watch tko boxing (2017)
- open win serena williams lead season return round th st (2018)

Using the different clusters to understand the different thematic for each year, I plotted the proportion of each cluster in each dataset (Figure 7). The bar plot underscores a discernible shift in the news coverage focus during a specific week in August for the years 2017 and 2018. Interest in Global Affairs & Social Issues surged in 2018, indicating a heightened focus on these topics, while the significant decrease in coverage of Cultural Events & Education suggests a waning interest or fewer events reported in that period. Crime & Legal Matters actually experienced an increase, hinting at either a rise in related events or a greater emphasis on legal reporting. Sports & Competition enjoyed a slight increase in coverage, potentially due to major events during that week. Meanwhile, Natural Disasters & Weather Events witnessed a substantial drop, suggesting fewer incidents or a shift in editorial focus away from such stories. The data from this particular week in August reflect a dynamic news environment with shifting priorities and public interest on a year-over-year basis.

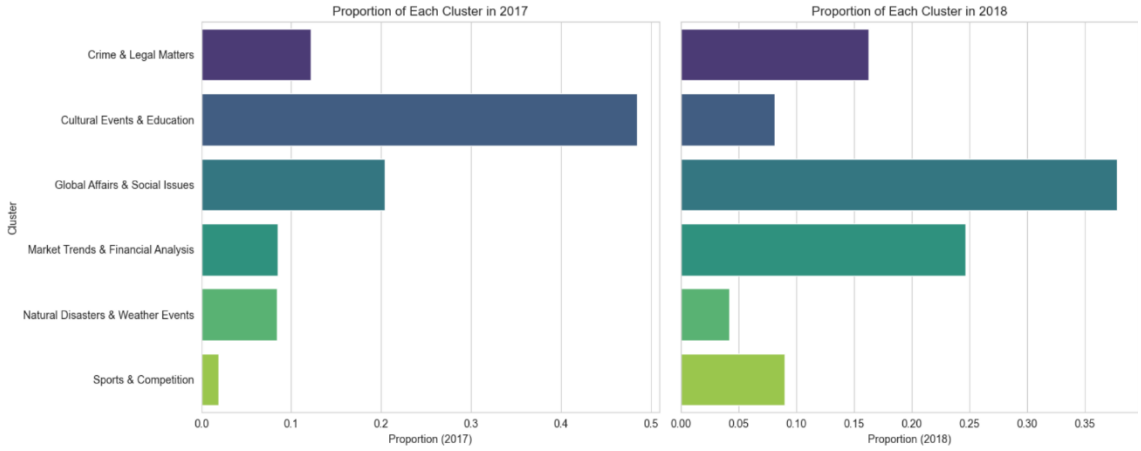


Figure 7 Clusters Proportion in headlines in 2017 and 2018

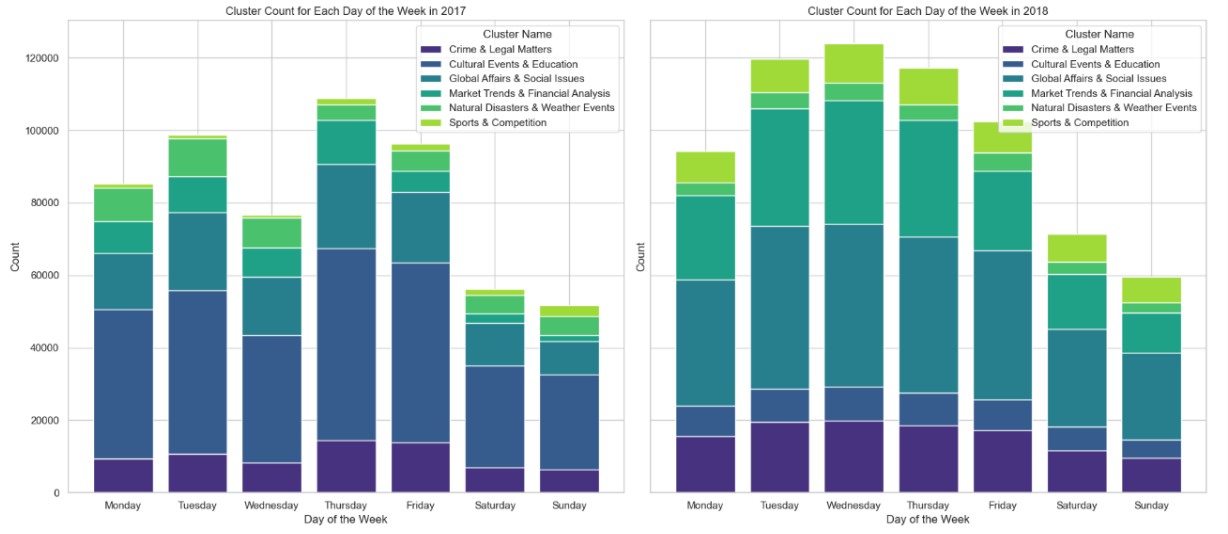


Figure 8 Cluster Count for each day in week of 2017 and 2018

From a different perspective, we can observe the amount of headlines published per team each day of the week. The stacked bar charts present the distribution of news topics for each day of the week in two different years, 2017 and 2018. Each cluster is represented by a different color, and the height of each colored segment indicates the count of news stories associated with that theme on a given day. In both years, there is a noticeable trend where certain topics have higher occurrences on specific days. For instance, in 2017, the category for Sports & Competition (represented in green) reaches its highest point on Sunday, aligning with the typical timing of numerous sporting events over the weekend. In contrast, in 2018, we do not observe this peak, suggesting an atypical spread of sports coverage throughout the week, possibly due to an increase in the number of sports events occurring during this particular week of the year or an unusual sport event. Market Trends & Financial Analysis (blue) have their lowest counts on weekends, reflecting the closure of financial markets. The distribution of other topics, such as Crime & Legal Matters and Natural Disasters & Weather Events, does not show a strong weekly pattern, indicating these news events occur more randomly throughout the week. Comparing the two years, some clusters show similar distributions on certain days, while others indicate shifts in focus or reporting frequency.

iii. Sentiment Analysis

In our examination, sentiment analysis was applied to decipher the emotional undertone of global news headlines across a designated week in 2017 and 2018. We utilized the DistilBERT model [2], a distilled version of the larger BERT architecture, which maintains most of the original model's performance on language understanding tasks but is more lightweight and faster. DistilBERT is the product of a process called knowledge distillation, where a compact model is trained to reproduce the behavior of a larger, more complex model. It achieves this by focusing on retaining the most informative aspects of the BERT model, which allows it to be efficient without a significant drop in accuracy. By using DistilBERT specifically for sentiment analysis, we were able to efficiently classify each headline into 'positive' or 'negative' categories. This provided a sentiment metric, affording us insights into the prevalent moods of global news during these intervals—highlighting whether the narratives were broadly viewed as optimistic or pessimistic.

Table 3 Positive and Negative Sentiment Count in 2017 and 2018

	2017	**2018**
Positive	231022	273360
Negative	342680	415174

The sentiment analysis from 2017 to 2018 presents intriguing findings. Initially, in 2017, the positive to negative sentiment ratio was 67%, but this slightly fell to 65.8% in 2018, hinting at a small rise in negative outlooks in late August 2018 news. Further investigation into specific themes shows that most have a ratio below 1 (as seen in Figure 9), signaling dominant negative sentiments and a generally pessimistic tone. Yet, in categories like Sports & Competition and Cultural Event & Education, the sentiment ratios exceed 1, with a notably more positive sentiment in 2018 despite the overall increase in negativity. Particularly in 2018, market trends and financial analysis reported one of the lowest ratios with Crime & legal matters reflecting a significantly pessimistic view in this sector and a growing negativity from the previous year. These changes in sentiment ratios highlight the complex emotional landscape reflected in news headlines over the two years.

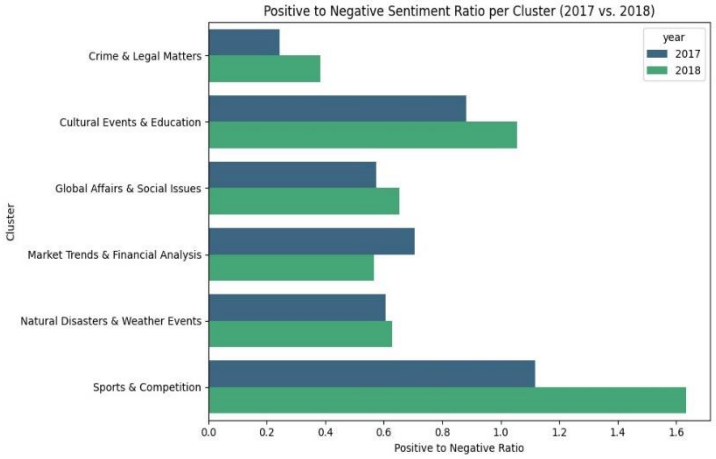


Figure 9 Positive to Negative Sentiment Ratio per cluster

The binary classification of sentiment often simplifies the complex nature of human emotions, as many articles can be neutral or have varying degrees of positivity and negativity. To delve into these subtleties, I explored two distinct sentiment analysis models: TextBlob [4] and the Sentiment Intensity Analyzer from the nltk.sentiment module. TextBlob is an accessible Python library offering API access to a range of Natural Language Processing (NLP) tasks, including sentiment analysis. It gives a polarity score indicating the text's overall sentiment. On the other hand, the Sentiment Intensity Analyzer (SIA), a component of the NLTK library, is a lexicon and rule-based sentiment analysis tool designed to discern the intensity of emotions in text. It provides a compound score reflecting the overall sentiment intensity, which is invaluable for identifying nuanced emotional expressions. Particularly noteworthy is that the SIA was initially trained in social media content. This origin makes it exceptionally suited for datasets like global news headlines, which often resemble the concise and impactful language typical of social media posts. Its training on varied and dynamic social media content potentially allows it to better understand and categorize the complex sentiments often found in the succinct and varied language of global news headlines. For the reasons mentioned, I chose to work with SIA for the sentiment analysis. The range of values that SIA provides is from -1 to 1. A score of -1 indicates extreme negativity, 0 represents neutrality, and 1 signifies extreme positivity. This range allows for a detailed and granular understanding of sentiment, capturing the subtle nuances between different degrees of positivity and negativity that are often present in textual data.

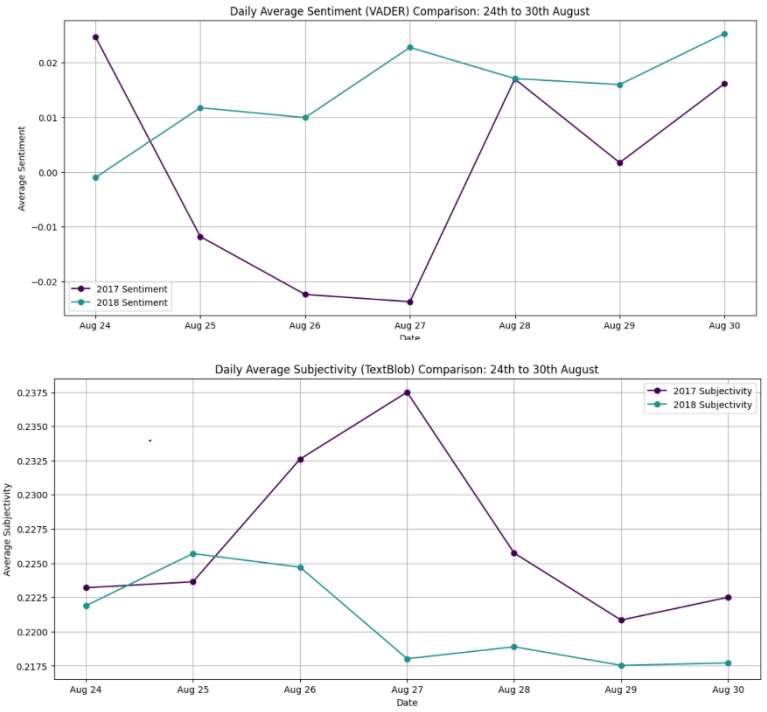


Figure 10 Sentiment and Subjectivity Time Series

However, I used TextBlob to measure subjectivity of headline news. In TextBlob, subjectivity is a measure used alongside polarity (sentiment) to provide a more nuanced analysis of textual data. While polarity ranges from -1 to 1, indicating the negativity or positivity of the text, subjectivity ranges from 0 to 1, where 0 is very objective and 1 is very subjective. Subjectivity in this context refers to the expression of personal feelings, opinions, or beliefs. A highly subjective text might express personal thoughts, feelings, or opinions, often found in personal blogs, opinion pieces, or art reviews. Conversely, a more objective text might present factual information, statistics, or other data that doesn't reflect a personal stance, such as a news report or a scientific article. In the Figure 10 above, we observe an inverse relationship between sentiment and subjectivity. This can occur because texts that are highly objective often refrain from exhibiting strong emotions or opinions, leading to a neutral sentiment score (close to 0). In contrast, subjective texts might express strong positive or negative sentiments, as they are likely to contain personal opinions or experiences. For instance, in a collection of news headlines, the more factual and unbiased the headline (high objectivity), the less likely it is to contain strong sentiment. On the other hand, opinion pieces or editorials (high subjectivity) are more likely to express strong sentiments, either positive or negative. Specifically, in 2017, we observed a notable decrease in sentiment polarity around the 27th of August, followed by an increase. This fluctuation might correspond to particular events or the nature of news coverage during those days. In contrast, 2018 presented a consistent increase in sentiment, demonstrating a steady rise in emotional expression in the news. However, the inverse occurred for subjectivity, highlighting the complex interplay between the factual reporting of events and the expression of opinions or sentiments within the news media. This nuanced relationship underscores the multifaceted nature of news reporting, where the tone and subjectivity of articles can significantly influence the perception and reception of news.

IV. Discussion

i. Global news context in last week of August of 2017 and 2018

To thoroughly understand the significance and implications of the data collected from August 4th to 30th, it's essential to consider the context of global events during this period. In 2017, Hurricane Harvey brought devastating destruction to Houston, contributing to a complex global political and security environment already tense from ongoing Middle Eastern conflicts and North Korea's nuclear threats. Transitioning to the same timeframe in 2018, the scene was marked by Senator John McCain's death, deeply impacting American political discourse. Concurrently, escalating trade tensions between the United States and China heightened fears of a trade war, affecting global market sentiments, and leading to significant stock market volatility. These events reflected the evolving geopolitical and economic uncertainties that influenced everything from local economies to international relations. Additionally, the sports scene in 2018 was dominated by the Asian Games and the US Open's intense final, where Naomi Osaka's victory over Serena Williams became a highlight. Understanding these events is critical for interpreting the dataset, as they provide context for the potential effects on global sentiment and economic trends. Recognizing the impact of these occurrences allows for a more nuanced analysis of the dataset's results.

ii. Results Interpretation and Reflection

Our descriptive and statistical analysis revealed an increase in the volume of articles in 2018 compared to 2017, indicating a more active news cycle. The longer headlines in 2018 suggested a trend towards more intricate reporting. This variance in news volume and detail provided the first layer of understanding of how media coverage evolved over the years. Temporal analysis showed consistent patterns in the daily and hourly distribution of news, with a notable surge on a specific day in 2018. Specifically, the temporal frequency analysis revealed a spike in coverage during the latter part of the week in 2018,

corresponding with the death of John McCain. The frequency analysis highlighted a shift from a focus on "Hurricane Harvey" and "Trump" in 2017 to a more varied set of themes including political figures and financial markets in 2018.

At the core of our analysis was the utilization of Non-negative Matrix Factorization for Topic Modeling, which unveiled distinct thematic transitions over the two years. In 2017, the predominant themes revolved around 'Hurricane Harvey and its catastrophic impact', 'President Trump's political maneuvers and pardoning of Joe Arpaio', as well as 'North Korea's missile launches and escalating tensions'. The attention was also drawn to 'the market forecasts', 'the finale of Game of Thrones', and 'Taylor Swift's music ventures'. These topics encapsulated the significant events and public preoccupations of the year. Conversely, 2018 presented a more varied thematic landscape. 'John McCain's memorial and legacy' and 'the Asian Games' took center stage, reflecting a shift in media narratives and public interest. Other prominent topics included 'global market forecasts', 'legal and criminal matters in various cities', and 'the tragic shooting in Jacksonville'. The year also saw discussions centered around 'trade deals and NAFTA negotiations', 'technological advances and public sector plans', and 'high-profile returns and performances in sports', particularly Serena Williams' comeback. The attention to 'pope's visit amidst the abuse scandal' and 'hurricane Lane's impact on Hawaii' marked the year's concern for social and natural issues. These thematic evolutions from 2017 to 2018, captured through Topic Modeling, illustrate not just the shifts in global events and narratives but also the changing priorities and interests of the public. From the focused discussions on politics and natural disasters in 2017 to a broader array of topics including sports, trade, and social issues in 2018, the media landscape mirrored the dynamic and multifaceted nature of global happenings and sentiments.

K-Means clustering corroborated these findings by organizing the topics into coherent clusters, showing a significant decrease in 'Natural Disasters & Weather Events' and an increase in 'Global Affairs & Social Issues' in 2018. The daily distribution of news themes provided further insights. Both years saw a decrease in financial news during weekends - a direct reflection of the financial markets' closure. However, 2017 experienced an uptick in sports coverage over the weekends, while 2018, with the Asian Games, had a consistent stream of sports news throughout the week. Moreover, 'Crime & Legal Matters' saw an increase in 2018, indicating a heightened focus and concerns on these issues.

Our Sentiment Analysis indicated a subtle shift towards a more pessimistic tone in news coverage from 67% positive in 2017 to 65.8% in 2018, particularly noted in the 'Market Trends & Financial Analysis' cluster. Additionally, an observed inverse relationship between sentiment and subjectivity in our data suggests that more objective texts tend to have neutral sentiment scores, while subjective ones exhibit stronger emotions. Specifically, in 2017, there was a notable fluctuation in sentiment polarity, with a decrease around August 27th followed by an increase, whereas 2018 showed a consistent rise in sentiment but an inverse trend for subjectivity, reflecting the complex dynamics between factual reporting and emotional expression in news media.

V. Conclusion

In fulfilling our objective to conduct an exhaustive exploratory data analysis along with advanced text analytics, our study has unearthed profound insights. Using Topic Modeling and Sentiment Analysis, we've effectively mapped out a detailed landscape illustrating how significant global events and prevailing narratives shaped the news during late August for 2017 and 2018. By contextualizing our findings against the backdrop of key occurrences - the tragic impact of Hurricane Harvey, the passing of John McCain, and the intensifying trade disputes - we've gained a nuanced perspective on their influence over media narratives. Our in-depth analysis underscores both the changes and consistencies in media emphasis and portrayal, delving into the intricate relationship between global events and their representation in news. Ultimately, our research has achieved its goal, offering an elaborate portrayal of how major worldwide events and prevailing stories shape news coverage during the highlighted weeks of August for two consecutive years. The discerned shifts and persistency's provide critical insights, enhancing our understanding of the dynamic nature of media focus and presentation, and mirroring the continually evolving fabric of global news narratives.

VI. References

[1] *Rohit Kulkarni (2018), One Week of Global Feeds [News CSV Dataset], doi:10.7910/DVN/ILAT5B, Retrieved from: Original paper by M Trampus, B Novak: Internals of An Aggregated Web News Feed*
Hosted By: Josef Stefan Institute, Slovenia: (http://ailab.ijs.si/si/people)
[2] distilbert-base-uncased-finetuned-sst-2-english · Hugging Face
[3] langdetect · PyPI
[4] https://textblob.readthedocs.io/en/dev/