

Project Report

Automated Competitor Analysis

IDS-572 – Deep Learning & Modern Application

Submitted by: Group

Abilash Sivakumar (UIN: 658587601)

Sruthi Srikanth (UIN: 670868541)

Khushi Saraswat (UIN: 674144424)



**College of Business Administration University of
Illinois at Chicago**

INTRODUCTION

The proliferation of textual data, particularly in the form of news articles and business reports, has increased dramatically in recent years. This surge presents both challenges and opportunities for deriving actionable insights from unstructured data. The focus of this project is on leveraging advanced data processing and analysis techniques to categorize and summarize news articles within the framework of SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats).

OBJECTIVES

This study aims to achieve the following objectives:

1. Develop efficient preprocessing pipeline for large-scale news article datasets using modern tools.
2. Classify the textual content into SWOT categories based on semantic context.
3. Generate concise summaries for each SWOT category to facilitate strategic decision-making.
4. Implement keyword-based filtering and visualization mechanisms to enhance data usability.

TOOLS AND TECHNOLOGIES

To meet these objectives, a carefully curated suite of tools and technologies was employed. Each tool was chosen for its specific utility in addressing various aspects of the project, as outlined below:

1. Data Handling and Preprocessing

- **Pandas:** Utilized for manipulating and analyzing structured data, enabling efficient handling of smaller dataset chunks.
- **Dask:** Leveraged for parallel and distributed processing to manage large datasets effectively.
- **Regular Expressions (Regex):** Employed for cleaning textual data and tokenizing sentences, facilitating structured extraction of meaningful information.

2. Natural Language Processing

- **Sentence Transformers:** Specifically, the all-MiniLM-L6-v2 model was used to generate context-aware sentence embeddings, providing accurate semantic representations.
- **Cosine Similarity:** Applied to measure similarity between sentence embeddings and predefined SWOT category vectors, ensuring accurate classification.

3. System and Computation

- **PyTorch (Torch):** Managed hardware resources, enabling GPU acceleration for efficient embedding generation.
- **TQDM:** Provided progress bars to monitor lengthy processes like batch embedding computation.

4. Key Integrations

- **Keyword-Based Filtering:** Incorporated user-defined keywords to filter articles by relevance.
- **SWOT Categorization Pipeline:** A systematic integration of preprocessing, analysis, categorization, and summarization tools.

DATASET OVERVIEW

The project utilized the **All the News 1.0** dataset, a comprehensive resource comprising diverse articles from multiple publishers. This dataset provided a rich base for SWOT categorization and summarization.

1. Dataset Structure

id: Unique identifier for each news article, **title:** The headline or title of the article, providing a concise summary of the content, **author:** The name of the journalist or author who wrote the article, **date:** The publication date of the article, **content:** The main body of the article, containing detailed information, **year and month:** Extracted fields representing the year and month of publication for temporal analysis, **publication:** The name of the publisher, useful for comparative analysis, **category:** Classification of the article, such as "Longform.", **digital:** Binary indicator, likely signifying whether the content is digital, **section:** Indicates a subsection of the publication, though it contains many missing values, **url:** The URL for the article, which also contains missing values.

2. Preprocessing Steps

The dataset was processed systematically to prepare it for analysis:

- **Data Extraction:** Retrieved and converted data from an SQLite database to a CSV file for easier manipulation.
- **Duplicate Removal:** Ensured unique entries by eliminating duplicates.
- **Handling Missing Values:** Filled missing titles and content with empty strings and removed rows with missing or invalid dates.
- **Whitespace Cleanup:** Removed unnecessary whitespace from text fields.
- **Boilerplate Removal:** Eliminated non-informative phrases (e.g., "Click here to read more").
- **Feature Engineering:**
 - Combined titles and content into a single "combined text" field, with higher weight assigned to titles.
 - Tokenized content into sentences for fine-grained semantic analysis.
 - Calculated word count for content richness assessment.
- **Datetime Parsing:** Converted date fields to datetime format and extracted year/month for trend analysis.

3. Dataset Characteristics

- **Diversity:** Articles from various publishers support comparative analysis.
- **Volume:** Large dataset suitable for temporal and categorical insights.
- **Content Variability:** Ranges from brief news pieces to detailed reports, requiring robust preprocessing and categorization methods.

4. Post-Processed Dataset:

The final dataset includes:

- A "combined text" field ready for embedding and categorization.
- SWOT category labels assigned to each article.
- Word counts to indicate article length.
- Cleaned and formatted fields, prepared for further analysis.

5. Challenges in Handling the Dataset

- **Missing Values:** Columns like sections and URLs contained many missing values and were not used in subsequent analyses.
- **Volume of Data:** The large dataset size required using Dask for efficient processing.
- **Noise in Text:** Articles often included non-informative or redundant phrases, which were removed during cleaning.

The image shown below is the dataset before cleaning:

id		title	author	date	content	year	month	publication	category	digital	section	url
0	1	Agent Cooper in Twin Peaks is the audience: on...	\nTasha Robinson\n	2017-05-31	And never more so than in Showtime's new...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN
1	2	AI, the humanity!	\nSam Byford\n	2017-05-30	AlphaGo's victory isn't a defeat for hum...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN
2	3	The Viral Machine	\nKaitlyn Tiffany\n	2017-05-25	Super Deluxe built a weird internet empi...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN
3	4	How Anker is beating Apple and Samsung at thei...	\nNick Statt\n	2017-05-22	Steven Yang quit his job at Google in th...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN
4	5	Tour Black Panther's reimagined homeland with ...	\nKwame Opam\n	2017-05-15	Ahead of Black Panther's 2018 theatrical...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN

The image shown below is the dataset after cleaning:

[36]:	date	title	content	publication
0	2017-05-31	Agent Cooper in Twin Peaks is the audience onc...	And never more so than in Showtimes new series...	Verge
1	2017-05-30	AI the humanity	AlphaGos victory isnt a defeat for humans its ...	Verge
2	2017-05-25	The Viral Machine	Super Deluxe built a weird internet empire Can...	Verge
3	2017-05-22	How Anker is beating Apple and Samsung at thei...	Steven Yang quit his job at Google in the summ...	Verge
4	2017-05-15	Tour Black Panthers reimagined homeland with T...	Ahead of Black Panthers 2018 theatrical releas...	Verge

MODELS AND TECHNIQUES

1. Sentence Transformers:

- **Model Used (all-MiniLM-L6-v2):**
 - A pre-trained transformer-based model for generating dense vector text representations. Built on Microsoft's MiniLM architecture, optimized for speed and memory efficiency.
 - **Purpose:** To create semantic embeddings for textual data, capturing contextual relationships between words and sentences.
- **Applications in the Project:**
 - **Embedding Generation:** Created embeddings for the combined_text column of the dataset.
 - Generated embeddings for predefined SWOT keywords.
 - **Semantic Similarity:** Compared article content to SWOT keywords using cosine similarity to determine the most relevant categories.
- **Why This Model?:** Efficient for large-scale text data processing. Produces high-quality embeddings for semantic analysis, ensuring accurate categorization.

2. Cosine Similarity

- **Purpose:**
 - Measures the cosine of the angle between two vectors, indicating the degree of similarity between them.
 - A score close to 1 implies high similarity, while a score near 0 indicates dissimilarity.
- **Applications in the Project:**
 - **SWOT Categorization:** Matched embeddings of article content with embeddings of SWOT keywords. Applied thresholds (e.g., 0.3 for sentences, 0.4 for combined text) to categorize articles.
 - **Keyword-Based Filtering:** Filtered articles based on their similarity to a user-defined keyword for targeted summarization.
- **Advantages:** Intuitive and computationally efficient. Effective in identifying relationships between high-dimensional text embeddings.

3. Regex-Based Sentence Tokenization

- **Purpose:** Splits the content into individual sentences to enable sentence-level semantic analysis.
- **Applications in the Project:**
 - **Key Sentence Extraction:** Tokenized content into sentences for embedding generation. Selected the top n most relevant sentences for summarization.
 - **Fine-Grained Categorization:** Enhanced SWOT classification by analyzing sentence-level semantic matches.
- **Why This Technique?:** Provides finer granularity in text analysis. Improves the quality of embeddings by focusing on smaller, meaningful chunks of text.

4. SWOT Categorization Pipeline

- **Process:**
 - **Predefined Keywords:** Each SWOT category was associated with a list of predefined

keywords (e.g., "market leader" for Strengths, "lawsuit" for Weaknesses).

- **Sentence-Level Analysis:** Calculated cosine similarity between sentence embeddings and SWOT keyword embeddings. Articles were categorized based on the maximum similarity score exceeding a threshold.
- **Combined Text Analysis:** Analyzed the overall embedding of combined_text against SWOT keyword embeddings to assign an overall category.
- **Iterative Refinement:** Articles initially categorized as "Null" were reprocessed with lower thresholds to improve coverage.

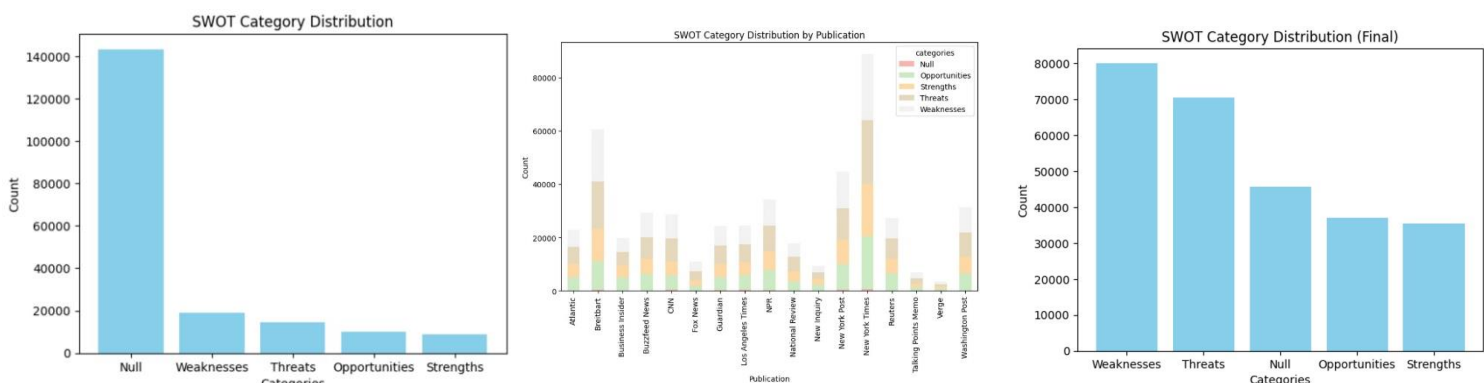
- **Creative Contributions:** Weighted title prioritization during embedding generation. Iterative refinement of null categories for improved accuracy.

5. Dask for Scalable Processing

- **Purpose:** Enabled memory-efficient handling of large datasets by splitting computations into smaller, parallel tasks.
- **Applications in the Project:**
 - **Data Loading:** Loaded the CSV file containing news articles in chunks, avoiding memory bottlenecks.
- **Embedding Computation:** Processed combined text in batches to compute embeddings for large datasets.
- **Why This Tool?:** Ideal for scaling Python-based computations to handle big data.

6. Challenges and Optimizations

- **Embedding Computation:**
 - **Challenge:** Generating embeddings for a large dataset was computationally intensive.
 - **Optimization:** Leveraged GPU acceleration via PyTorch and processed embeddings in batches of 1000 articles, balancing speed and memory efficiency.
- **Categorization Ambiguity:**
 - **Challenge:** Many articles were initially categorized as "Null" due to unclear alignment with SWOT categories.
 - **Optimization:** Refined thresholds iteratively, lowering them gradually to capture borderline cases and improve classification coverage.
- **Recategorization of Null Entries:**
 - **Threshold Adjustment:** Articles initially labeled as "Null" were re-evaluated with reduced thresholds (sentence-level: 0.2, combined-text: 0.3) to improve classification coverage.
 - **Updated Categorization:** Similarity scores were recalculated, and categories were reassigned wherever possible. Articles that still lacked sufficient alignment retained the **Null** label.



- **Semantic Noise:**
 - **Challenge:** Irrelevant or vague content led to misclassification and poor-quality embeddings.
 - **Optimization:** Filtered out articles under 100 words and removed boilerplate text during preprocessing to improve data quality.

7. Inferences from the Models

1. Strengths:

- SentenceTransformers generated high-quality embeddings, enabling robust semantic categorization.
- Cosine similarity offered efficient, scalable semantic alignment for large datasets.

2. Limitations:

- Static keywords occasionally missed nuances in complex articles, highlighting the need for dynamic or context-aware keyword sets.
- Lower thresholds for "Null" reclassification improved coverage but sometimes led to borderline categorization, suggesting a need for adaptive thresholding.

SUMMARIZATION PIPELINE

The summarization pipeline was designed to condense vast amounts of text into actionable insights while aligning content with predefined SWOT categories. It was built with a focus on scalability and precision, leveraging embeddings and filtering techniques to deliver relevant summaries

1. Key Steps in the Pipeline

• Loading the Dataset:

- Given the large size of the dataset, Dask was used to load the CSV file efficiently, enabling chunk-wise processing without exhausting system memory.
- Missing values in the title and content columns were replaced with empty strings to ensure consistent processing.

• Embedding Generation:

- The combined text column, which merges article titles and content, was passed through the SentenceTransformer model to generate contextual embeddings.
- Embeddings were computed in batches of 1000 articles to optimize memory usage and computational speed, with GPU acceleration further improving efficiency.

• Keyword-Based Filtering:

- Articles were filtered based on their semantic similarity to user-defined keywords. For example, if the keyword "growth" was specified, the cosine similarity of each article's embedding was computed against the embedding of the keyword.
- A similarity threshold of 0.5 ensured that only highly relevant articles were retained for summarization.

• Extracting Key Sentences:

- Tokenized the filtered articles into sentences using regex to break down the content into smaller, meaningful chunks.
- Extracted up to five key sentences from each article based on relevance, ensuring that the summary provided a concise yet comprehensive view.

• SWOT-Based Summarization:

- Articles were grouped by their assigned SWOT categories, such as Strengths or Weaknesses.
- Summaries were created by concatenating the most relevant sentences for each category, emphasizing readability and precision.
- If a category lacked sufficient content, a placeholder such as "No relevant content" was added.

• Visualization:

- A bar chart was created to visualize the number of articles per publisher post-filtering. This helped understand the distribution of summarized content across publications.

2. Strengths of the Pipeline

- **Scalability:** Using Dask and batching allowed the pipeline to handle large datasets efficiently without compromising performance.
- **Relevance:** Keyword-based filtering and cosine similarity ensured that only the most pertinent articles were summarized.

- **Clarity:** Extracting key sentences resulted in summaries that were concise, easy to read, and actionable.

3. Areas for Improvement

- **Dynamic Thresholding:** The similarity threshold was static, which may have excluded borderline-relevant content. A dynamic approach based on content variability could improve flexibility.
- **Enhanced Keyword Matching:** Incorporating synonyms or related terms dynamically using embeddings could capture a broader range of relevant content.
- **Category Overlap:** Articles are sometimes aligned with multiple categories (e.g., Strengths and Opportunities). Refining the scoring mechanism could improve category precision.

RESULTS AND INSIGHTS

The analysis and summarization processes yielded valuable insights into the dataset, revealing trends and patterns across SWOT categories and publishers. This section highlights the key results and discusses their implications.

1. Key Results

- **SWOT Categorization:**
 - The articles were successfully classified into four SWOT categories (Strengths, Weaknesses, Opportunities, and Threats) based on semantic similarity:
 1. **Strengths:** Included content emphasizing competitive advantages like market leadership and strong financial performance.
 2. **Weaknesses:** Highlighted challenges such as product recalls, lawsuits, and operational inefficiencies.
 3. **Opportunities:** Focused on growth prospects, new markets, and technological innovations.
 4. **Threats:** Identified risks such as economic downturns, competition, and regulatory challenges.
 - **Category Distribution:**
 1. Strengths and Opportunities had the highest representation, indicating that articles often highlight positive business aspects.
 2. Weaknesses and Threats were comparatively less frequent but provided critical insights.
- **Keyword-Based Filtering:**
 - Filtering articles with specific keywords like "growth" or "profit" helped extract highly targeted insights.
 - The pipeline effectively narrowed the dataset to relevant content without overwhelming users with irrelevant articles.
- **Publisher Trends:**
 - Certain publishers dominated specific SWOT categories. For instance:
 1. **Strengths** were more frequently reported by business-oriented publications.
 2. **Threats** were often highlighted in general news outlets.
 - Visualizations revealed these trends clearly, helping identify the focus areas of each publication.
- **Summarization:**
 - Key sentences extracted from articles provided concise and actionable summaries for each SWOT category.
 - For example:
 1. **Strengths:** "The company continues to dominate market share with an innovative product portfolio."
 2. **Weaknesses:** "The recall has impacted brand reputation, leading to a decline in customer trust."

2. Insights

- **Content Trends:**
 - Business articles predominantly focused on Strengths and Opportunities, aligning with their target audience's interests in growth and success.
 - Weaknesses and Threats, while less frequent, offered critical risk analyses that are valuable for strategic planning.
- **Keyword Effectiveness:**
 - Semantic filtering based on keywords like "growth" and "risk" demonstrated how targeted insights could be extracted. This makes the pipeline highly adaptable to user-specific queries.
- **Publisher Focus:**
 - Publications had distinct areas of focus, suggesting that integrating these insights could provide a comprehensive SWOT analysis for businesses operating in diverse industries.

3. Challenges and Learnings

1. **Category Overlap:**
 - Some articles aligned with multiple categories (e.g., an opportunity might also be a threat depending on perspective). Refining the scoring mechanism could reduce overlaps.
2. **Reclassification:**
 - Articles classified as "Null" in the initial round required reprocessing with adjusted thresholds. While this improved coverage, a dynamic categorization model could further enhance accuracy.
3. **Summarization Limitations:**
 - Extracting only five key sentences from large articles occasionally missed important context. Expanding the number of sentences or using an abstractive summarization model could improve coverage.

4. Visual Highlights

1. **Category Distribution:**
 - A bar chart showing the number of articles in each SWOT category revealed a higher focus on Strengths and Opportunities compared to Weaknesses and Threats.
2. **Publisher Trends:**
 - A stacked bar chart illustrated how different publishers contributed to each SWOT category, highlighting their editorial focus.
3. **Keyword Trends:**
 - Line graphs showed temporal trends in keyword mentions, providing insights into how certain topics evolved over time.

The output of our model:

Strengths Summary:

The iPhone came first but it was of secondary importance. Apple's big event this month relegated the company's best known and best selling product to the role of an opening act for the next chapter in Apple's history. The Watch. With a perfectly pitched tremor of emotion in his voice, CEO Tim Cook finally confirmed all the rumors of an Apple wearable device while leaving some mystery about how it will work and everything it will do. The product itself was almost less important than the statement the com

Weaknesses Summary:

SAN FRANCISCO With sales sluggish and stiffening competition from rivals like Google and Facebook, Apple announced on Monday coming improvements to the software that runs its devices, including a revamped Music app, an easier login process and better information sharing across devices. More significantly, it also made it easier for app developers to tap the full power of Apple's technology from its Siri voice assistant to instant messaging and Apple Pay in their own apps. We believe it is crucial that w

Opportunities Summary:

The iPhone came first but it was of secondary importance. Apple's big event this month relegated the company's best known and best selling product to the role of an opening act for the next chapter in Apple's history. The Watch. With a perfectly pitched tremor of emotion in his voice, CEO Tim Cook finally confirmed

all the rumors of an Apple wearable device while leaving some mystery about how it will work and everything it will do. The product itself was almost less important than the statement the company made.

Threats Summary:

Apple management hosted a dreary worldwide launch of the iPhone 7 where the company's managers heaped praise on themselves for having the courage to get rid of the earphone jack. The best applause at Apple's glitzy September 7 shindig at the Bill Graham Civic Auditorium in Downtown San Francisco came in response to news that Pokémon Go will now be available for play on the updated Apple Watch 2. From a business standpoint, the few thousand 12-year-olds that can afford a 400 watch must undoubtedly be excited.

The below is the output from ChatGPT for the prompt 'go through the news article between 2013 to early 2018 and provide a concise and short SWOT analysis on apple.inc'

Strengths

Apple Inc. demonstrated unparalleled innovation and market leadership between 2013 and early 2018. The successful launches of flagship products like the iPhone 6/6S, iPhone X, and Apple Watch reinforced its dominance in the premium consumer electronics segment. The company's ecosystem, which seamlessly integrated hardware, software, and services, fostered exceptional brand loyalty and customer retention. Financially, Apple remained a powerhouse, achieving consistent revenue growth, high profit margins, and amassing substantial cash reserves, giving it the ability to invest heavily in research and development.

Weaknesses

Despite its successes, Apple was heavily reliant on iPhone sales, which accounted for approximately 60% of its revenue during this period. This dependency made the company vulnerable to market saturation and potential declines in smartphone sales. Additionally, the premium pricing of Apple products limited its ability to penetrate emerging markets and appeal to budget-conscious consumers. Furthermore, Apple faced challenges in its supply chain, relying on third-party manufacturers like Foxconn, which subjected the company to periodic scrutiny over labor practices and production delays.

Opportunities

Emerging markets such as India and China presented significant opportunities for Apple, where growing demand for smartphones and premium devices remained largely untapped. The company's expanding services segment, including Apple Music, iCloud, and the App Store, showed strong growth potential, offering a steady stream of recurring revenue. Furthermore, advancements in augmented reality (AR) and artificial intelligence (AI), exemplified by features like Face ID, provided Apple with new avenues for innovation and market expansion.

Threats

Apple faced intense competition during this period from rivals such as Samsung, Huawei, and Google, which aggressively pursued market share with innovative and competitively priced products. Regulatory scrutiny also emerged as a significant threat, particularly in the European Union, where the company faced tax investigations and antitrust challenges. Additionally, the smartphone market showed signs of saturation in developed economies, potentially limiting growth prospects and necessitating diversification into other product categories and markets.

This analysis highlights Apple's strategic position during this period, showcasing its strengths and opportunities while acknowledging the vulnerabilities and threats it needed to address.

Similarly for Copilot

Between 2013 and early 2018, Apple Inc. demonstrated significant strengths, including strong brand loyalty, consistent financial performance, and a reputation for innovation with products like the Apple Watch and advancements in iPhone technology. However, the company faced weaknesses such as a heavy dependence on iPhone sales and high product prices, which could limit market share in price-sensitive regions. Opportunities for Apple included expanding into emerging markets like China and India, and continuing to innovate in new product lines such as wearables and services. On the flip side, Apple faced threats from intense competition with other tech giants like Samsung and Google, as well as increasing regulatory challenges in various markets. This period highlighted Apple's ability to leverage its strengths while navigating its weaknesses and external threats.

Conclusion:

while the project successfully demonstrated the potential of NLP techniques for SWOT categorization and summarization, its performance was constrained by limited computational resources and a relatively small dataset. These limitations impacted the ability to experiment with advanced models and optimize categorization thresholds for greater precision. Despite these challenges, the project established a scalable pipeline and showcased the feasibility of using embeddings and similarity metrics to extract actionable insights from unstructured data. Future work should focus on leveraging larger datasets, enhancing computational capacity, and integrating more dynamic, context-aware approaches to improve accuracy and scalability.

APPENDIX

Citation:

- <https://chatgpt.com/share/67513dcd-e8d8-8003-be08-7b088ce5eaea>

EDA

Preview of News Dataset: The output displays the first five rows of the loaded news dataset, confirming successful loading. It includes columns such as date, title, article, and publication.

```
[16]: df.head()
```

		id	title	author	date	content	year	month	publication	category	digital	section	url
0	1	Agent Cooper in Twin Peaks is the audience: on...	\nTasha Robinson\n	2017-05-31	And never more so than in Showtime's new...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN	
1	2	AI, the humanity!	\nSam Byford\n	2017-05-30	AlphaGo's victory isn't a defeat for hum...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN	
2	3	The Viral Machine	\nKaitlyn Tiffany\n	2017-05-25	Super Deluxe built a weird internet empi...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN	
3	4	How Anker is beating Apple and Samsung at thei...	\nNick Statt\n	2017-05-22	Steven Yang quit his job at Google in th...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN	
4	5	Tour Black Panther's reimagined homeland with ...	\nKwame Opam\n	2017-05-15	Ahead of Black Panther's 2018 theatrical...	2017.0	5.0	Verge	Longform	1.0	NaN	NaN	

Statistical Description of Data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179309 entries, 0 to 179308
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   date        179309 non-null  datetime64[ns]
 1   title       178764 non-null  object
 2   content     177191 non-null  object
 3   publication 179309 non-null  object
dtypes: datetime64[ns](1), object(3)
memory usage: 5.5+ MB

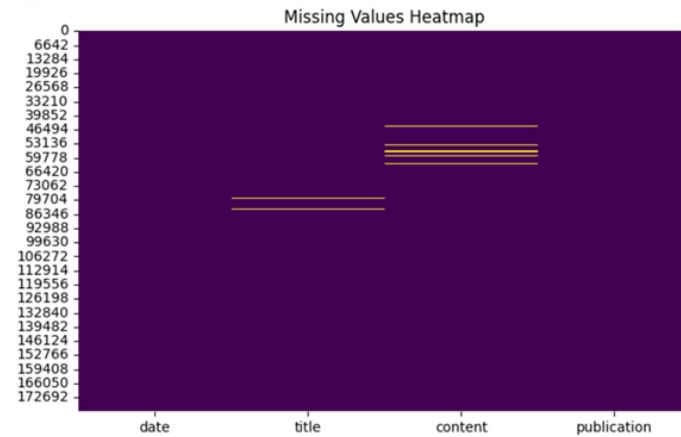
None

count          date          title \
unique          NaN          178764
top            NaN          Sunday Reading
freq           NaN          250
mean    2016-09-28 20:28:31.450066432    NaN
min          2000-05-15 00:00:00    NaN
25%          2016-05-12 00:00:00    NaN
50%          2016-11-10 00:00:00    NaN
75%          2017-03-29 00:00:00    NaN
max          2018-03-31 00:00:00    NaN

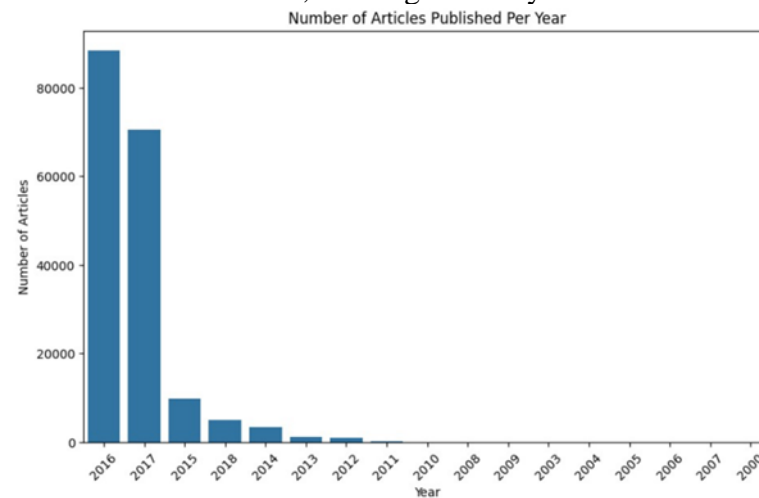
count          content          publication
unique          177191          179309
top    Astrophysicistxa0and StarTalk Radio hostxa0xa0...    New York Times
freq          83          30141
mean          NaN          NaN
min          NaN          NaN
25%          NaN          NaN
50%          NaN          NaN
75%          NaN          NaN
max          NaN          NaN
```

Missing Values Analysis: This heatmap illustrates the percentage of missing values in the dataset's key columns. While the date and publication columns have no missing values, the title and content columns show 0.3% and 1.18% missing values, respectively.

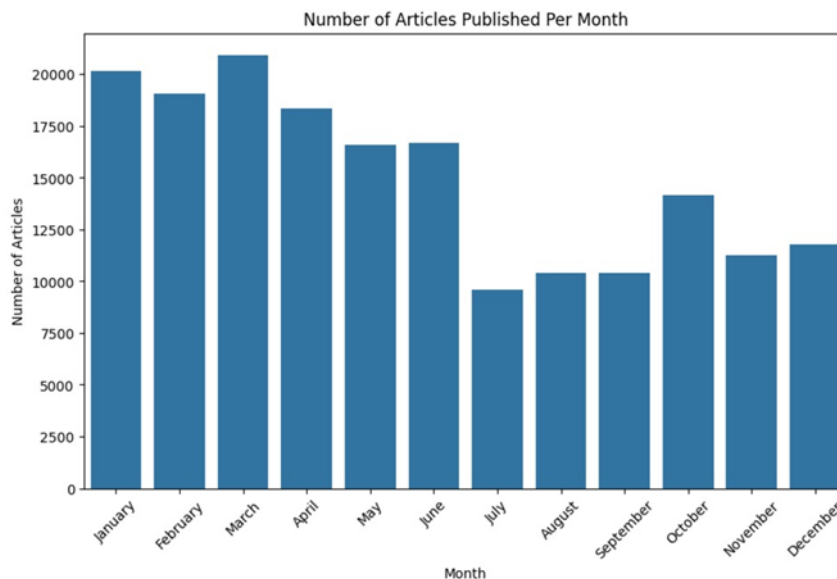
```
Missing Values (%):
date      0.000000
title     0.383945
content   1.181201
publication 0.000000
dtype: float64
```



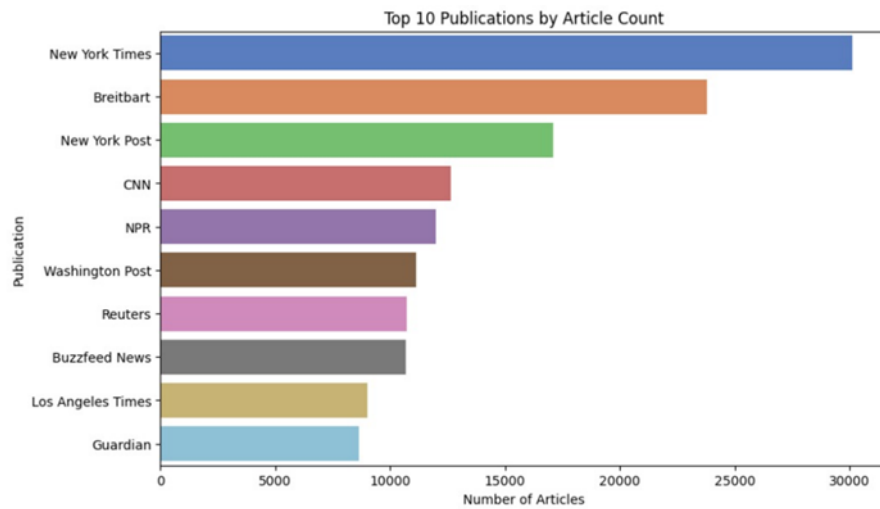
Number of Articles Published Per Year: This bar chart displays the number of articles published per year. The dataset is heavily skewed toward 2016 and 2017, with significantly fewer articles in earlier years.



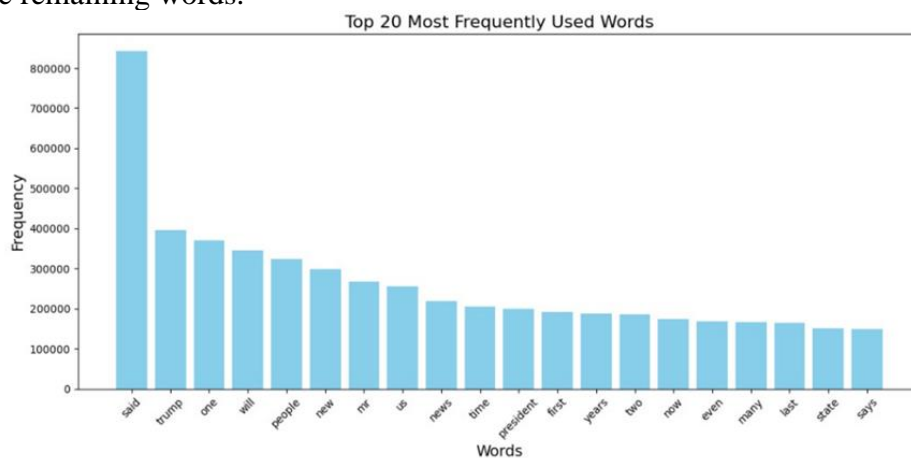
Monthly Trends in Article Publications: This bar chart shows the number of articles published per month. March has the highest publication count, followed by January and February, while July and August have the lowest counts.



Top 10 Publications by Article Count: The bar chart shows the top 10 publications by the number of articles published. The New York Times leads with the highest count, followed by Breitbart and the New York Post. Other notable publications include CNN, NPR, and the Guardian.



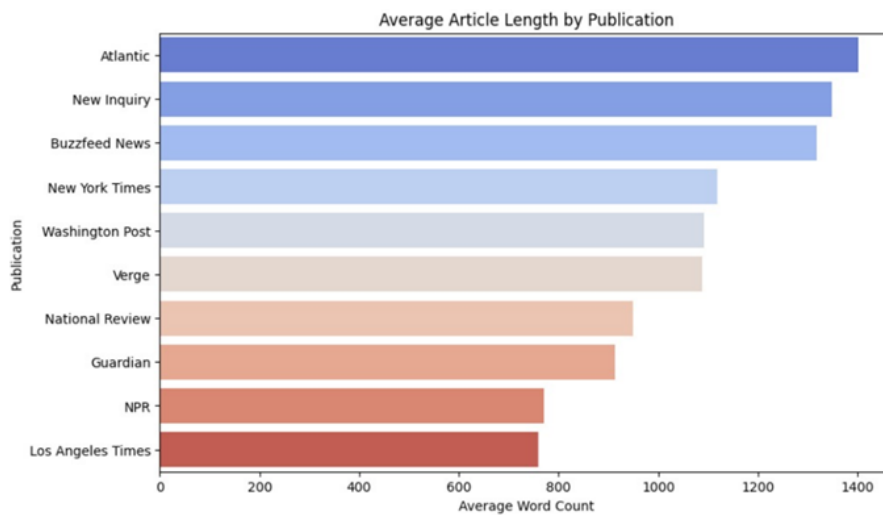
Top 20 Most Frequent Words in Text Data: The bar chart displays the top 20 most frequently used words, with "said" being the most common, followed by "trump," "one," and "will." The frequency decreases steadily across the remaining words.



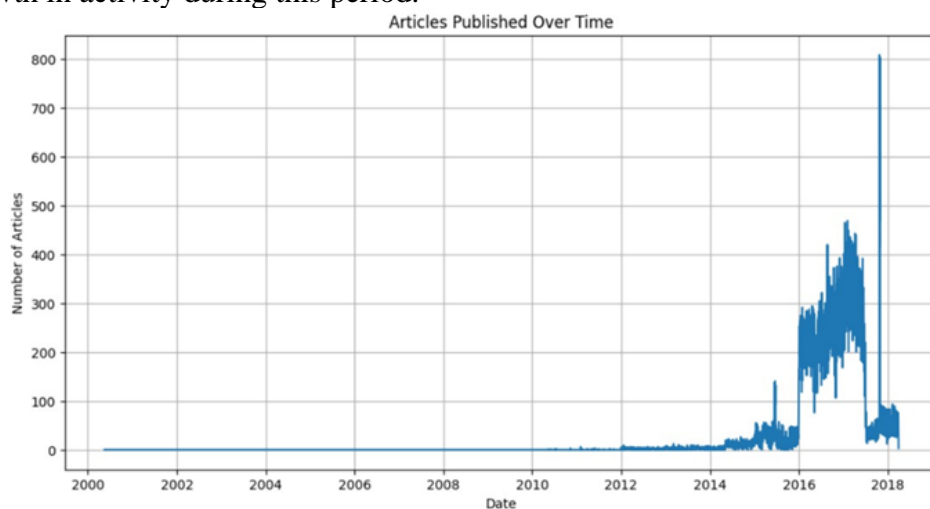
Word Cloud of Article Content: The word cloud visualizes the most frequent words in article content, with "said," "trump," "people," and "news" being the most prominent. Other notable terms include "president," "new," and "like."



Average Article Length by Publication: The chart compares average article lengths across various publications, with "The Atlantic" having the highest average word count and "Los Angeles Times" among the lowest. It highlights differences in editorial style and depth.



Trends in Article Publications Over Time: The graph shows the number of articles published over time, with a sharp rise starting around 2013, peaking in 2017, and declining slightly afterward. This indicates a significant growth in activity during this period.



Correlation Between Year and Word Count: The heatmap shows a strong self-correlation for both variables (1.00) and minimal correlation (0.04) between publication year and word count, indicating little relationship between the two.

