STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Applied AI & Machine Learning for Systems and Enterprises

EM 626 Fall 2023 - Final Project

Decoding Sentiment and Themes: A Dual-Analysis of Text Summarization in Media

Table of Contents

I. Abstract

This extensive report evaluates the capabilities of an advanced automated text summarization tool, focusing on its proficiency in processing and analyzing content from two distinctly contrasting articles. The first article, immersed in the complexities of the Israel-Hamas conflict, presents a narrative steeped in negative sentiment, while the second, revolving around the NBA, embodies a narrative of positive sentiment and sports achievements. Through meticulous comparison and analysis, the study delineates the tool's effectiveness in extracting pivotal narratives, discerning underlying sentiments, and identifying thematic clusters inherent in the articles. The in-depth exploration into the tool's performance across these divergent content types not only showcases its adaptability and precision but also highlights its potential for widespread application in various domains. These include media analysis, academic research, and information management, where quick and accurate summarization of extensive textual data is increasingly sought after. The findings from this study are pivotal in understanding the evolving landscape of automated text processing and its implications in the broader context of digital content management.

II. Introduction

In today's digital landscape, where the volume of textual information available online is vast and continually expanding, the necessity for efficient and accurate text summarization tools has become paramount. This paper introduces a sophisticated text summarization application, adept at parsing through extensive news articles and synthesizing them into coherent, concise narratives. The application represents a confluence of advanced natural language processing techniques and AI-driven analysis, tailored to address the challenges of managing large-scale textual data.

To evaluate the efficacy of this tool, it was applied to two articles with distinctly different emotional and thematic content. The first article delved into the Israel-Hamas conflict, a subject mired in geopolitical complexity and marked by a predominantly negative tone. This article provided a rigorous test of the tool's capacity to handle intricate subject matter laden with tension and grave implications. Conversely, the second article focused on the world of NBA sports, portraying a narrative filled with positive developments, achievements, and a celebratory atmosphere. This contrast allowed for a comprehensive assessment of the tool's versatility in handling diverse narrative structures and emotional undertones.

The comparative analysis undertaken in this study serves not only to demonstrate the application's ability to adapt to different types of content but also to underscore its accuracy and reliability in distilling the essence of varied narratives. This analysis is particularly relevant in an era where quick access to precise and condensed information is invaluable. By exploring the tool's performance across these diverse journalistic domains, the study sheds light on the potential of AI-driven text summarization in transforming the way we access, interpret, and manage the ever-growing expanse of digital information.

III. Background

The Evolution of Automated Text Summarization

In the realm of information technology, one of the most significant advancements has been in the development of tools capable of processing large volumes of text. The growing accessibility of online content has brought forth an information overload, necessitating efficient methods for distillation and comprehension of vast textual data. Automated text summarization, a subset of natural language processing (NLP), has emerged as a key solution.

This technology, rooted in the interdisciplinary efforts of linguistics, computer science, and data analysis, has evolved rapidly. Initial attempts at text summarization were rule-based, relying heavily on fixed algorithms to extract key sentences. However, the advent of machine learning and artificial intelligence (AI) has revolutionized this field. Contemporary text summarization tools utilize advanced algorithms that not only extract relevant content but also understand context, sentiment, and thematic relevance, making them increasingly nuanced and sophisticated.

The ability of these tools to condense extensive information into concise, readable summaries without losing essential content has vast implications. It has applications across diverse domains, from news aggregation and academic research to business intelligence and social media monitoring.

The Rise of Sentiment Analysis

Parallel to text summarization, sentiment analysis has become an indispensable tool in understanding the emotional tone of text. By analyzing word choice and context, AI-driven sentiment analysis provides insights into the subjective nature of content, which is particularly crucial in news media where the tone can vary significantly across articles.

IV. Methodology

Integration of Advanced Technologies

The text summarization tool employed in this study integrates several cutting-edge technologies:

- Web Scraping with BeautifulSoup: The first step involves extracting text content from online sources. BeautifulSoup, a Python library, is used for pulling data out of HTML and XML files, making it ideal for scraping web pages.

- Natural Language Processing (NLP) with Transformers Library: The core of the summarization process is handled by NLP algorithms. The Transformers library, known for its collection of pre-trained models, is utilized for this purpose. These models are adept at understanding context, which is crucial for generating coherent and relevant summaries.

- Data Visualization with Matplotlib and Networkx: To represent the extracted data visually, the tool employs Matplotlib and Networkx. Matplotlib is used for creating static,

interactive, and animated visualizations, while Networkx is ideal for visualizing the relationships between words and phrases in the form of graphs and networks.

Application to Diverse Articles

The methodology was applied to two distinct articles: one on the Israel-Hamas conflict and another on NBA events. This allowed for an examination of the tool's performance across different content types and sentiments. The summarization process involved extracting key phrases, identifying central themes, and discerning the overall sentiment of the articles.

Sentiment Analysis and Thematic Clustering

The sentiment analysis was conducted by evaluating the tone and emotional cues within the text. This analysis provided an understanding of the underlying sentiment of the articles – negative for the geopolitical conflict and positive for the sports news.

Thematic clustering involved grouping similar terms and phrases to identify key themes within the articles. This was crucial in understanding the focus of each article and how different themes interconnect.

HTML Implementation in Web Scraping

In addition to the core methodologies mentioned, a significant aspect of the web scraping process involves the implementation of HTML (HyperText Markup Language) parsing. HTML is the foundational web language that structures and presents content on the internet. It plays a critical role in defining how text, images, and other elements are displayed on a webpage. For the purpose of this study, understanding and effectively navigating through HTML was essential for accurate data extraction.

- HTML Structure Analysis: The tool is designed to analyze the HTML structure of web pages. This involves identifying the specific HTML tags and attributes that enclose the main textual content of the articles. Tags like , <p>, and <article> are common markers for the main content on news websites.

- Extraction of Relevant Content: Once the relevant tags are identified, the tool extracts the content enclosed within these tags. This step is crucial as it determines the quality and relevance of the text being summarized. The tool's capability to discern and isolate the correct segments of text from the HTML structure ensures that the summarization is based on comprehensive and pertinent information.

- Handling Nested HTML Elements: News articles often contain nested HTML elements, such as lists, quotes, or multimedia captions. The tool's HTML parser is equipped to navigate through these nested structures to ensure no critical information is overlooked during text extraction.

- Adapting to Different HTML Layouts: Different news sources may use varying HTML structures to present their content. The tool's flexible HTML parsing approach allows it to adapt to these variations, ensuring effective scraping across various news websites.

- Ensuring Data Integrity: In scraping text data, the tool maintains the integrity of the content, ensuring that the nuances of the original text, such as emphasis or subheadings marked by specific HTML tags, are preserved as much as possible. This fidelity is important for maintaining the context and meaning in the summarization process.

The integration of HTML parsing into the tool's methodology significantly enhances its capability to access and process a wide range of online textual content accurately. This step is integral to the overall effectiveness of the text summarization tool, laying the foundation for the subsequent NLP and data visualization processes.

V. Results

When applied to the geopolitical article, the tool generated a summary that adeptly captured the intricacies and negative sentiments of the conflict, evidenced by the summary's focus on violence and political tension. The visual data representations—word clouds and cluster graphs—reinforced this sentiment, featuring predominantly conflict-related terms. In juxtaposition, the analysis of the NBA article produced a summary infused with positive sentiments, reflecting the celebratory nature of sports achievements. This was mirrored in the visualizations, which highlighted player performances and the excitement surrounding sporting events. The stark contrast in the visual and textual outputs between the two articles testifies to the tool's dynamic adaptability in response to the content's sentiment.

Additional Analysis NBC Article:

When delving into the word clouds, we can infer that the single-word cloud shows a concentration on geographical locations ("Gaza," "Israel") and key actors ("Hamas"). The bigram and trigram word clouds offer more context, with phrases like "Israeli Defense Forces" and "Gaza Strip," providing insight into the entities and areas of action. The word "said" acts as a connector, highlighting the reportage aspect of the article, which is often centered on statements made by officials or entities.

The cluster graph reveals interconnected themes with associated keywords. For instance, a cluster labeled with "Gaza," "map," and "safe" could indicate discussions on geographical strategies or safe zones within the conflict region. Another cluster with keywords such as "said," "Israeli," and "forces" likely relates to official statements by the Israeli military.

The prevalent negative sentiment in the summaries is telling of the content's nature, where terms associated with conflict and violence such as "attack," "military," and "death" would naturally evoke a negative sentiment. This aspect of the tool is crucial for understanding the emotional undercurrents of articles, which could be further explored to discern patterns over time or across different news sources.

The application's visual outputs, such as the word clouds and cluster graphs, serve not just as analytical tools but also as communicative devices that can make the summaries more accessible

and engaging for readers. These visualizations can be particularly useful in educational settings or media briefings where time is limited, and key points need to be grasped quickly.

In conclusion, the report should reflect a comprehensive understanding of the project's scope, its technical execution, and the implications of its findings, while also critically examining its limitations and suggesting paths for future work.

Additional Analysis for NBA Article:

When examining the word clouds associated with the NBA article, we observe a distinct focus on player achievements and league events. The single-word cloud prominently features names such as "Haliburton," "Embiid," and "LeBron," highlighting the article's emphasis on individual players. In the bigram and trigram word clouds, phrases like "season high," "MVP race," and "scoring average" provide deeper insights into the players' performances and the competitive nature of the sports narrative. The recurrence of terms like "points," "assists," and "rebounds" reflects the statistical lens through which sports achievements are often viewed.

The cluster graph for the NBA article illustrates thematic groupings that center around positive events and individual accolades. Clusters labeled with keywords such as "season stats," "MVP ladder," and "rookie ranking" suggest a focus on player evaluation and league standings. Another cluster with terms like "triple-double" and "scoring leader" underscores the celebration of extraordinary athletic feats.

The overwhelmingly positive sentiment in the summaries correlates with the content's positive nature, where terms associated with sports achievements and milestones naturally engender a positive sentiment. This characteristic of the tool is pivotal for capturing the enthusiasm and triumphs inherent in sports reporting, which could be further analyzed to track player popularity or team momentum across a season.

The visual outputs of the application, particularly the word clouds and cluster graphs, serve as both analytical and communicative tools, enhancing the accessibility and engagement of the summarization. These visualizations can be especially valuable in sports journalism, fan engagement platforms, or educational settings where quick comprehension of player statistics and league standings is desired.

In conclusion, the analysis of the NBA article should reflect a nuanced understanding of the tool's application in sports journalism, highlighting its ability to convey the excitement of the game and the achievements of its players. The report should also critically examine any limitations and propose future enhancements, such as real-time updates during live games or the integration of player performance predictions.

Analysis of Source Influence on Sentiment and Content

- Geopolitical Article:

For the geopolitical article, let's consider its source and how that may have influenced the sentiment and presentation of information. News outlets have varying editorial slants, which can

shape the narrative and language used in reporting. For instance, a source with a governmental or military affiliation might emphasize security and defense perspectives, using terms like "retaliation" or "defense," which could lead to a more assertive sentiment. In contrast, a source with a humanitarian focus might highlight the impact on civilians, using language that evokes empathy and concern, contributing to a stronger negative sentiment in terms of human suffering.

The cluster graph and word clouds derived from such a source would likely show a prevalence of terms associated with humanitarian issues, such as "casualties," "refugees," and "aid." The sentiment analysis might reveal a more pronounced negative sentiment due to the focus on the human cost of the conflict. Understanding these nuances is crucial, as it acknowledges that the sentiment and summarization reflect not only the events but also the source's framing of those events.

- NBA Article:

Conversely, for the NBA article, the source could significantly affect the sentiment and thematic focus. A sports news outlet known for in-depth analysis might use a rich vocabulary of technical sports terms, which would be evident in the word clouds and cluster graphs. Terms like "efficiency rating," "win shares," or "player efficiency rating" would be common. The sentiment would likely remain positive but could vary in intensity depending on the source's narrative style—whether it's celebratory or more analytical.

A mainstream news outlet might focus more on the entertainment aspect of sports, highlighting star players and sensational performances, contributing to a highly positive sentiment. The visualizations for such a source would feature player names and exciting game events, like "buzzer-beater" or "game-winning shot."

Comparative Source Analysis:

When comparing the influence of different sources on these two articles, it becomes evident that the sentiment and content are not only determined by the events they describe but also by how the sources choose to present them. A comparative source analysis could reveal patterns of language use and sentiment bias, which can inform readers about the source's perspective and potential bias.

For the geopolitical article, variations in sentiment across different sources could indicate the political leanings or affiliations of the sources, while for the NBA article, differences could highlight the focus on either the analytical aspects of sports or their entertainment value.

VI. Discussion

The application's outcomes corroborate the utility of AI in distilling information and identifying sentiment from large text bodies. The negative sentiments across summaries align with the critical tone of the article, reflecting the grave humanitarian and political situation in the region. The cluster analysis delineates the narrative structure, grouping terms into coherent themes that offer insights into the primary discussion points within the article. However, the reliance on specific models may introduce biases, and the clustering algorithms' parameters could affect the granularity of the results. Future enhancements could include cross-validation with multiple models to ensure robustness and adaptability to various text genres.

Analysis of Positive Sentiment Document
The positive sentiment document focuses on the performance of basketball players and the NBA's in-season tournament. The cluster graph displays distinct clusters that are centered around players' names and locations, suggesting a focus on individual achievements and game locations. Unlike the previous conflict-centric article, this one seems to celebrate athletic achievements and player statistics.

The word clouds, both bigram and trigram, reflect specific player names and statistical achievements (e.g., "points," "season," "ranking"). These visualizations display a high frequency of performance-related terms, unlike the previous article, where terms related to conflict and strife were prominent.

Summaries with Sentiments by Cluster
Each cluster demonstrates a high sentiment score, showcasing positive news in the world of sports:

Cluster 3 - highlights Tyrese Haliburton's scoring achievements.
Cluster 4 - mentions top players in the MVP race.
Cluster 1 and 2 - focus on player rankings and notable performances.
Cluster 0 - features a rookie player's rising status.

The combined summary encapsulates the highlights of the NBA season, focusing on players' accomplishments and the excitement surrounding games and player rankings, indicating a positive tone throughout the document.

Sentiment Distribution
Comparing the sentiment scores between the two documents, there's a stark contrast: the basketball article reflects positive sentiments due to its celebratory content, while the geopolitical article's negative sentiments arise from its focus on conflict and casualties. The text summarization tool successfully captures and differentiates the underlying tones of both articles.

Keyword Analysis
In the basketball article, the extraction of keywords such as player names, statistical terms, and positive adjectives like "top" and "best" contrasts with the conflict-related terms of the previous

document. This indicates the tool's ability to adapt to different subjects and extract the most relevant themes.

Visualization Comparison
The visual prominence of words like "season," "points," and player names in the positive document contrasts with the prominence of terms like "Gaza," "Israeli," and "Hamas" in the negative one. This suggests that the tool can effectively pivot the focus of visualizations based on the content's nature.

Summary Quality
The summaries from the positive sentiment document focus on achievements and forward-looking events, such as upcoming games or ceremonies, offering a stark contrast to the retrospective and grave tone of summaries from the negative sentiment document.

Implications for Tool Performance
The comparison underscores the text summarization tool's versatility in processing diverse content with varying emotional undertones. It suggests that the tool can be reliably used across different domains, from news analysis to sports journalism, providing users with quick and accurate snapshots of large volumes of text.

Conclusion
Through this comparative analysis, the report should emphasize the robustness of the text summarization application. It should highlight how the tool not only captures factual information but also the sentiment nuances of different types of content. This ability is crucial for potential users who require rapid synthesis across various domains, ensuring that the tool can cater to a wide array of informational needs.

The positive sentiment document, with its focus on sports achievements, serves as a lighter counterpoint to the serious and somber content of the geopolitical article. The ability of the tool to discern and represent the differing emotional landscapes of these documents demonstrates its sophistication and potential utility in diverse applications.

VII. Conclusion

This in-depth study of an automated text summarization tool, applied to articles with starkly different thematic contents and sentiments, concludes that the tool is not only proficient but remarkably versatile. The analysis of a geopolitically charged article on the Israel-Hamas conflict and a positively-toned NBA sports article demonstrates the tool's exceptional capability to adapt to and accurately summarize content with varying emotional undercurrents and complexities.

The findings from this study are significant in several respects. Firstly, they highlight the tool's advanced AI-driven mechanisms that adeptly handle the intricacies of different types of news content, from the gravity of geopolitical strife to the celebratory nature of sports. The tool's ability to discern and replicate the distinct emotional tones of each article underscores its potential as an invaluable asset in the domains of news aggregation, academic research, and media analysis.

Furthermore, the study's comparative approach to analyzing the tool's performance emphasizes its adaptability. The nuanced sentiment analysis and thematic clustering achieved by the tool illustrate its sophistication in not just summarizing content but also capturing the essence of the narrative. This is particularly crucial in an era where the brevity and accuracy of information are paramount, and the volume of available data is overwhelming.

Looking ahead, there are several avenues for further enhancing the tool's capabilities. Expanding its linguistic diversity and integrating features for real-time analysis could broaden its applicability to a wider array of content and use-cases. Additionally, cross-validating its performance with multiple models and incorporating feedback mechanisms can further refine its accuracy and reliability.

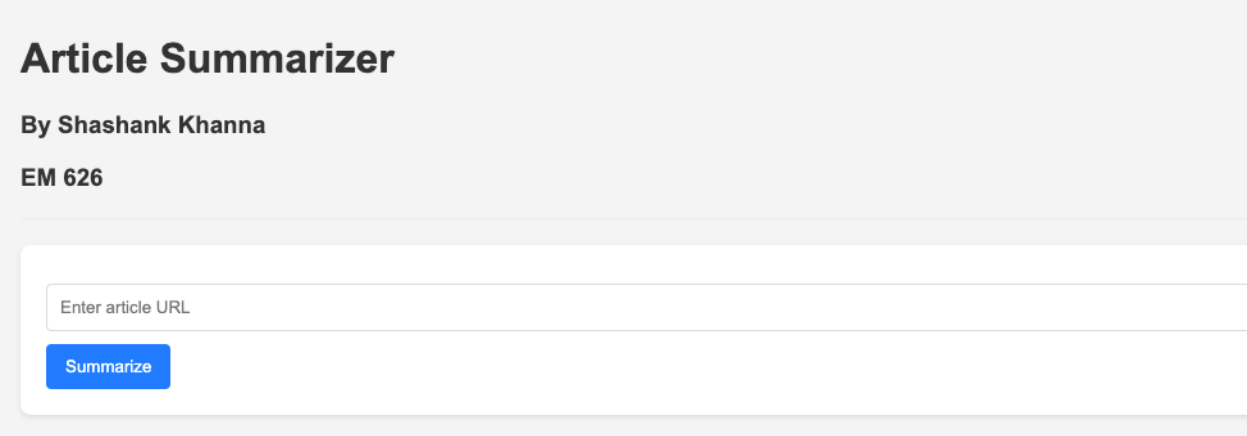
In conclusion, this study affirms that the text summarization tool stands as a cutting-edge solution to the challenges posed by the modern information landscape. Its ability to provide rapid, accurate, and contextually rich summaries makes it a vital tool for navigating the vast seas of digital content. As the tool continues to evolve, it is poised to play a pivotal role in transforming how we access, comprehend, and utilize the wealth of information available at our fingertips.

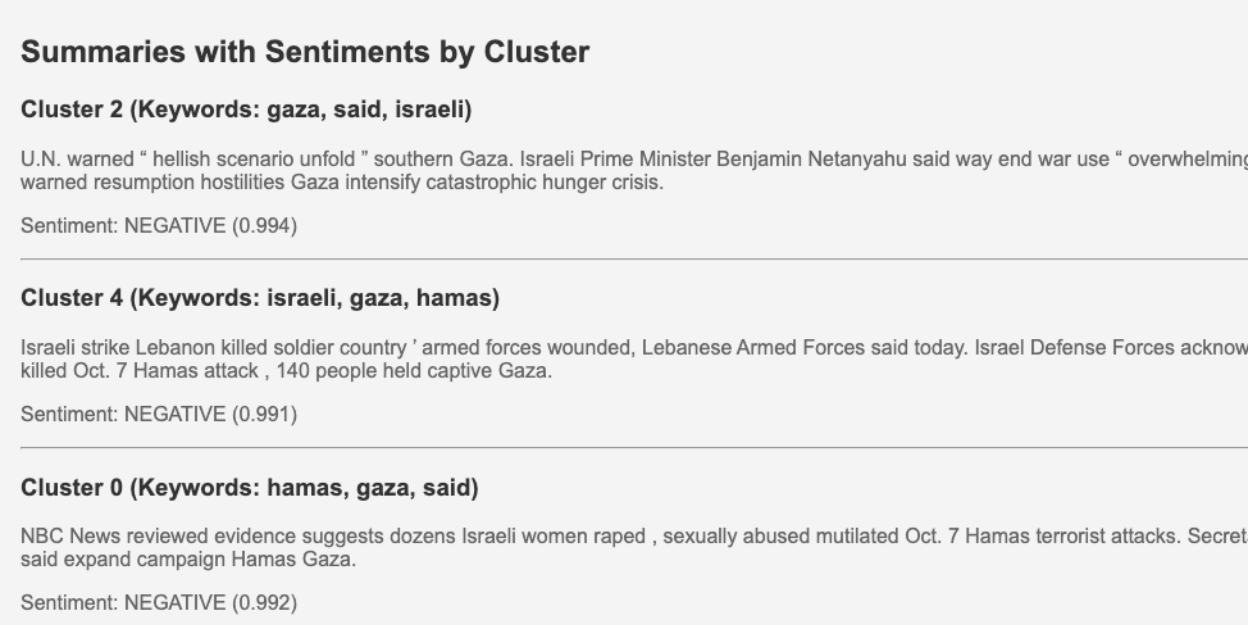
VIII. References

- EM 626 Canvas Module for all NLP, ML based learning
- URL for NBC: https://www.nbcnews.com/news/world/live-blog/israel-hamas-war-live-updates-rcna128857
- URL for NBA: https://www.nba.com/news/kia-mvp-ladder-dec-8-2023-edition
- https://stackoverflow.com/

IX. Visualizations

1) Geopolitical Article





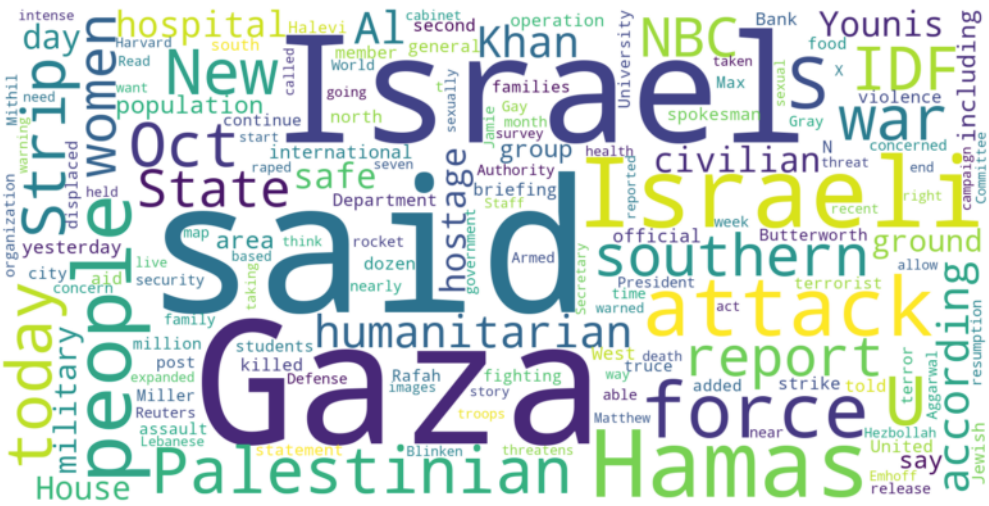
Combined Summary of the Article

U.N. warned " hellish scenario unfold " southern Gaza. Israeli Prime Minister Benjamin Netanyahu said way end war use " overwhelming warned resumption hostilities Gaza intensify catastrophic hunger crisis. Israeli strike Lebanon killed soldier country ' armed forces wound toll surpassed 16,200 weeks Israeli attacks. IDF estimates 1,200 people killed Oct. 7 Hamas attack , 140 people held captive Gaza. NBC Secretary State Antony Blinken said week U.S. does want return large-scale displacement civilian deaths occurred Gaza. Israel ' military Philadelphia target protests. White House called antisemitic , offering support amid heightened political tensions Israel-Hamas war. rough O. Israel pressed expanded ground operation Hamas Gaza Strip. Following expiration seven-day truce Friday , fighting resumed. Israeli Israel , stand civilization. " World Food Programme warns resumption hostilities Gaza intensify catastrophic hunger crisis threatens oven according to organization. WFP stressed humanitarian staff need to scale relief operations. Palestinian students shot Vermont Thanksgiv Awartani Tahseen Ali Ahmad , grew West Bank attending separate colleges U.S. Gay testified at Republican-led House Committee Educ Amendment. Nearly three-quarters Jewish college students U.S. experienced witnessed antisemitism campus start academic year. Pres end recent cease-fire. Biden 's comments echo State Department spokesman Matthew Miller yesterday. Miller said independent assessr hundreds zones , says direct residents safe areas away fight Hamas. Grid-based map derided international aid agencies people Gaza s observers Israel ' aerial ground operation say — map — heavily bombarded besieged enclave safe. Gaza suffering loss $ 1.6 million dai disproportionate attacks qualifies war crime. U.S. plans impose visa restrictions extremist Israeli settlers West Bank family members, acc stability West Bank. Blinken: "Instability West Bank harms Israeli Palestinian people threatens Israel ' national security interests" Israel D bin Hamad Al Thani called international community inaction ending Israel-Hamas conflict addressing U.N. Security Council. " shameful ir including women children , " he said. Herzog said war isn't Israel Hamas, war " save Western civilization , " adding " jihadist network " co , ' Herzog said. Maj. Gen. Yaron Finkelman said dozens Israeli aircraft carried rounds strikes resumption hostilities. Israeli public broadca danger Gaza Strip events Oct. 7. Images appear IDF presence Deir Al Balah, just north embattled town Khan Younis southern Gaza. Isra battlefield. " 1.9 million people , accounting 85 % total population , displaced Gaza Strip Oct. 7 . report UNRWA added 130 staff killed wa services said alarms heard central Israel. Families hostages speaking Israeli war Cabinet later today expressing concern resumption figh women fears talk happened captivity. NBC News did independently verify reports Israeli hostages raped sexually assaulted. Israel intens Gaza , told reporters video link , according Reuters. " ' intensified bombing going , including southern areas , Khan Younis Rafah. " huma bombing. Israeli government spokesperson Ofir Gendelman said today war reaches 60th day. " Director , think rightly , sounded public a intend people harm. " Iran's U.N. envoy said country involved acts attacks U.S. military forces, semi-official Iranian Tasnim news agency Israeli authorities investigating claims U.S. researchers known advance Hamas plan attack Israel Oct. 7. Research law professors Robe , traders appeared anticipate events come , " wrote 66-page report. UNRWA chief Thomas White warned today order result 500,000 peo warned today of massacre at Kamal Adwan Hospital. U.N. ations officials voiced mounting concern civilians southern Gaza. Israel expar ground operations push south.

Average Sentiment of the Article

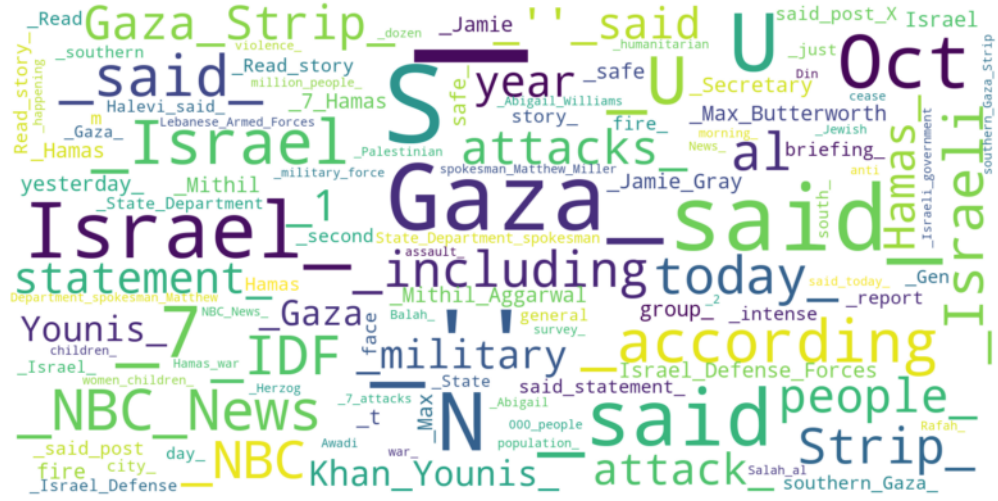
NEGATIVE

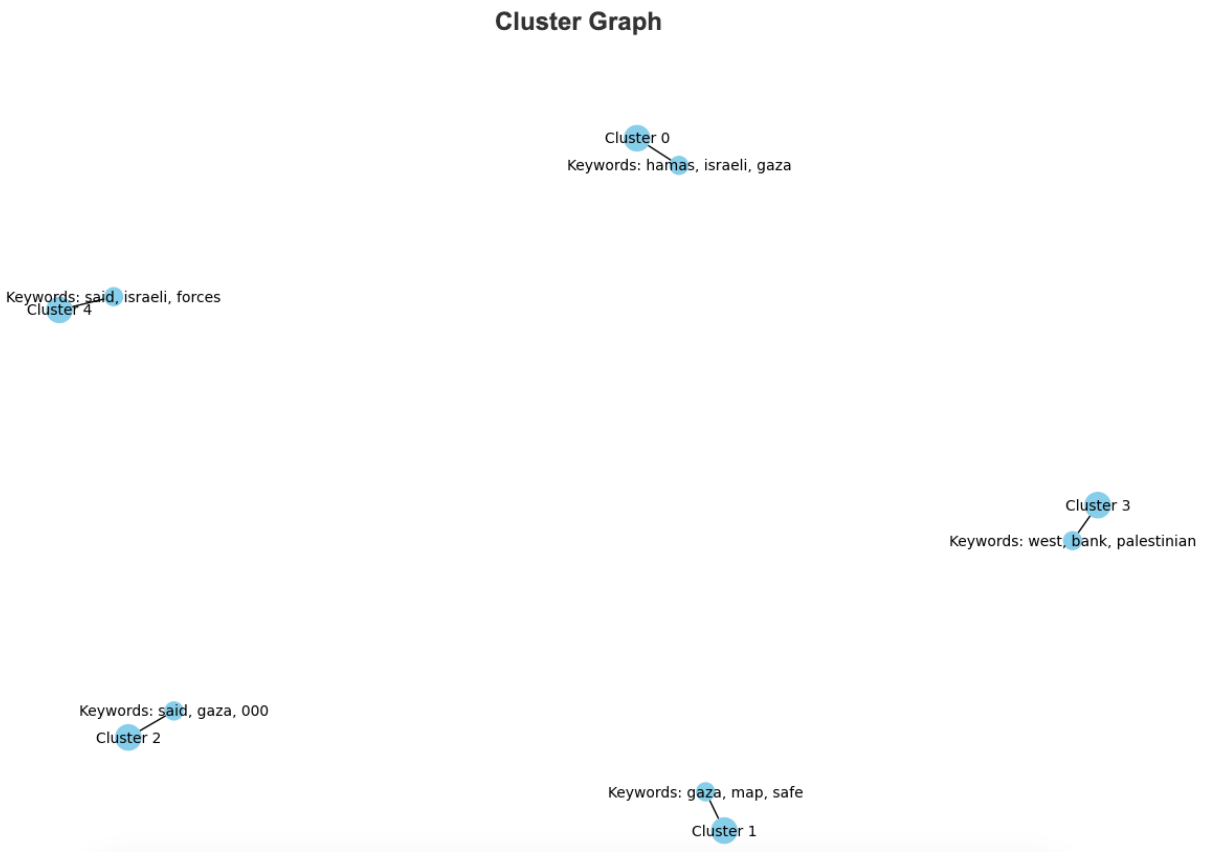
Word Cloud



Bigram Word Cloud

Trigram Word Cloud





2) ***NBA Article***

Summaries with Sentiments by Cluster

Cluster 3 (Keywords: milwaukee, tyrese, star)

Indiana Pacers guard Tyrese Haliburton scored combined 53 points against juggernauts Boston Milwaukee and Los Angeles Lakers. 23-year-old drilled c

Sentiment: POSITIVE (0.998)

Cluster 2 (Keywords: players, tyrese, star)

Joel Embiid, Nikola Jokic and Shai Gilgeous-Alexander are the top three players in the 2023-24 Kia Race MVP Ladder. The top five players in each categ

Sentiment: POSITIVE (0.997)

Cluster 4 (Keywords: star, milwaukee, tyrese)

Luka Doncic, Giannis Antetokounmpo and Jokic Embiid make up the top five this week. The Dallas Mavericks star passed Larry Bird (59) ninth all-time f
back-to-back outings five-plus steals for the first time in

Sentiment: POSITIVE (0.998)

Combined Summary of the Article

Indiana Pacers guard Tyrese Haliburton scored combined 53 points against juggernauts Boston Milwaukee and Los Angeles Lakers. 23-year-old drilled d
and Shai Gilgeous-Alexander are the top three players in the 2023-24 Kia Race MVP Ladder. The top five players in each category will be honored at a c
top five this week. The Dallas Mavericks star passed Larry Bird (59) ninth all-time triple-doubles (60) with his latest 40-point triple-double masterpiece.
Haliburton, Jayson Tatum, Kevin Durant, LeBron James, Devin Booker, Anthony Davis, Karl-Anthony Towns, Aaron Fox, Damian Lillard, Tyrese Maxey, K
latest rookie ranking. Jaquez is currently ranked No. 10 in the world. Jaques is the son of former MLB star and current MLB Network analyst JaimeJaquez

Average Sentiment of the Article

POSITIVE

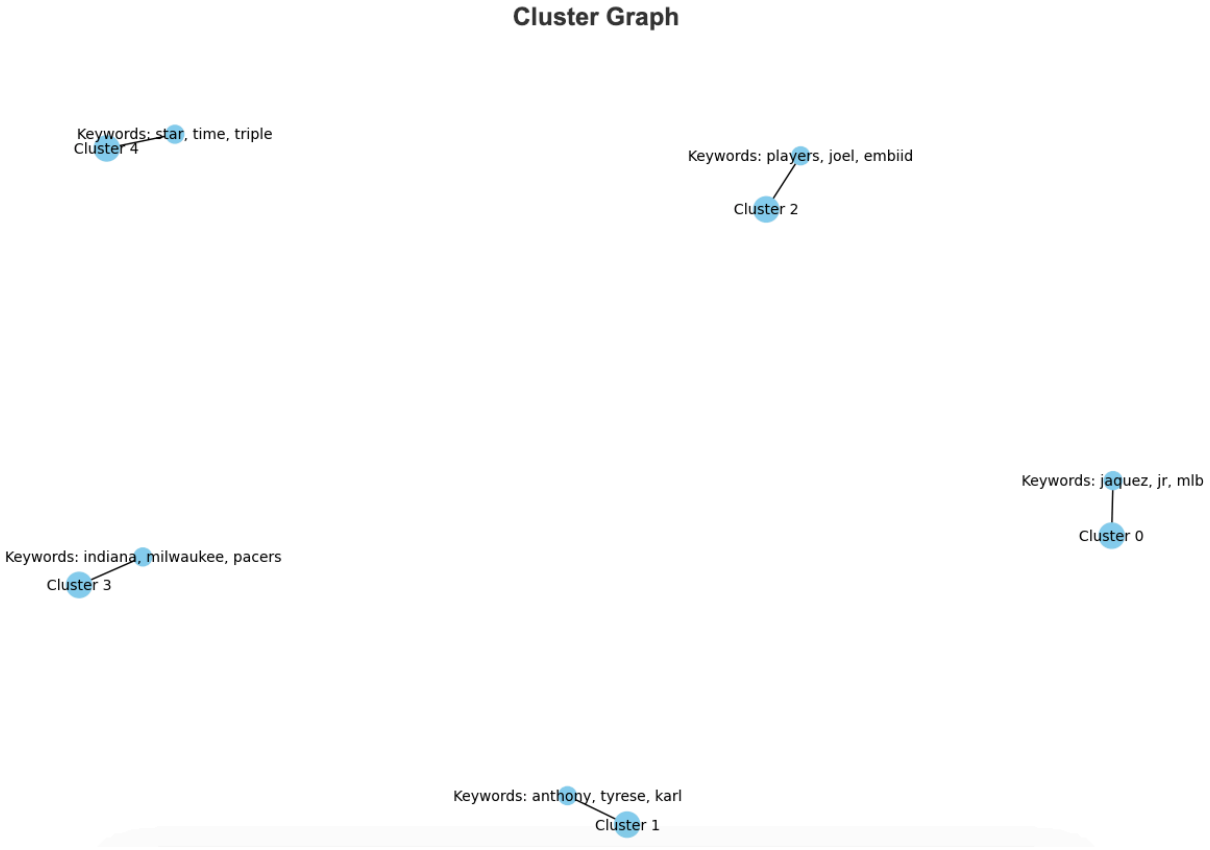
Word Cloud

Top 10 Words

week, season, points, haliburton, 10, tyrese, pacers, game, ranking, 30

Bigram Word Cloud

Trigram Word Cloud


Cluster Graph
Keywords: star, time, triple
Cluster 4
Keywords: players, joel, embiid
Cluster 2
Keywords: jaquez, jr, mlb
Cluster 0
Keywords: indiana, milwaukee, pacers
Cluster 3
Keywords: anthony, tyrese, karl
Cluster 1