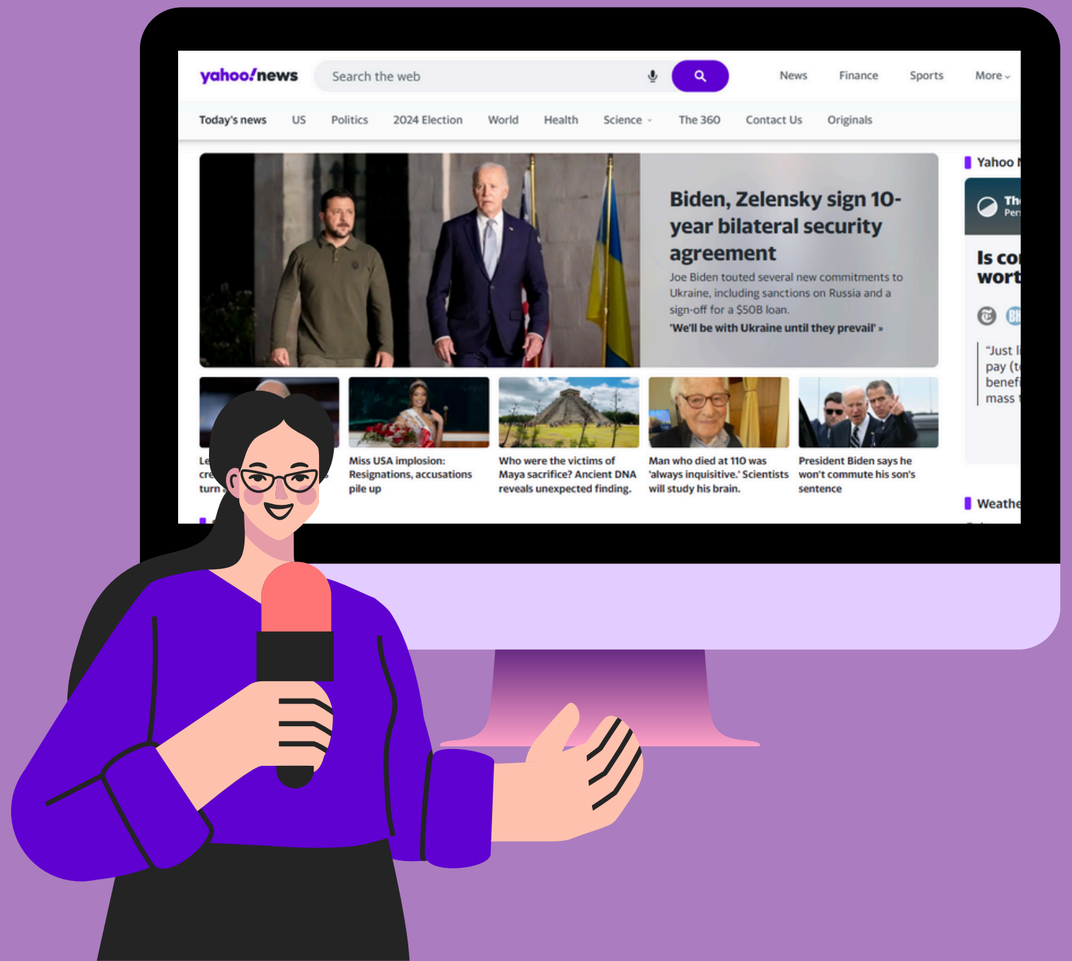
PROJECT SCRAPPING

Meisyatul Ilma - G1501231073

yahoo!news



Find the latest and most reliable news from your device with Yahoo News!

As one of the world's leading news portals, Yahoo News brings you the latest information with high journalistic quality. With in-depth coverage, incisive analysis, and exclusive news, we ensure you always get accurate and relevant information. Yahoo News delivers the news you need anytime and anywhere, from politics and economics to entertainment and technology. Enjoy easy access and stay up-to-date with important events from around the world. Yahoo News - your trusted source of information!

COLLECTING DATA YAHOO!NEWS

This research uses web scraping techniques from the news.yahoo.com site to collect data effectively and accurately. Using the R programming language, data consisting of headlines, publication dates, and news article links are stored in MongoDB. The scraping process is automatically set to collect one piece of data daily, from May 29, 2024, to June 13, 2024, ensuring a consistently updated dataset.

The scraping process was automated to collect one data every day, starting from May 29, 2024, to June 13, 2024, to ensure a consistent and up-to-date dataset. The research also used GitHub as a version control, collaboration, backup, documentation, and automation tool, which made the scraping process more efficient, reliable, and manageable. With this approach, the research collected rich and relevant data continuously.

Data Processing

Connecting to MongoDB Atlas

Data Cleaning

Stopwords

Word Cloud Visualization

Keyword Extraction

Tokenization and Data Frame Creation

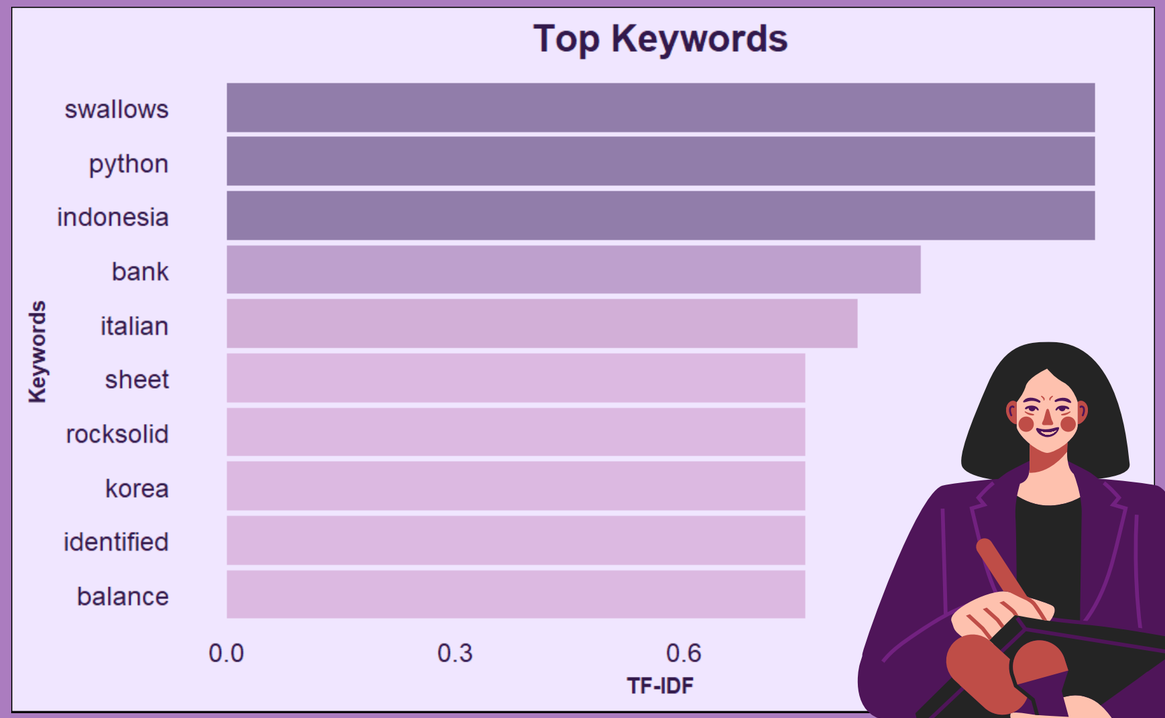
Topic Modeling with LDA

Visualization of LDA Topics

Sentiment Analysis

Keyword Extaction

The results of the keyword extraction analysis using the TF-IDF algorithm show that keywords such as Swallows, Python, and Indonesia stand out with the highest TF-IDF values, indicating the importance of these topics in the analyzed document set. Swallows may refer to specific references in the text, while Python most likely relates to a popular programming language. Indonesia suggests a lot of information or discussion related to this country. Other keywords such as Bank, Italian, and Korea also emerged as significant, reflecting important economic, cultural, and geographical topics. This analysis helps identify and understand the main topics and trends in the corpus in a more in-depth and informative manner.



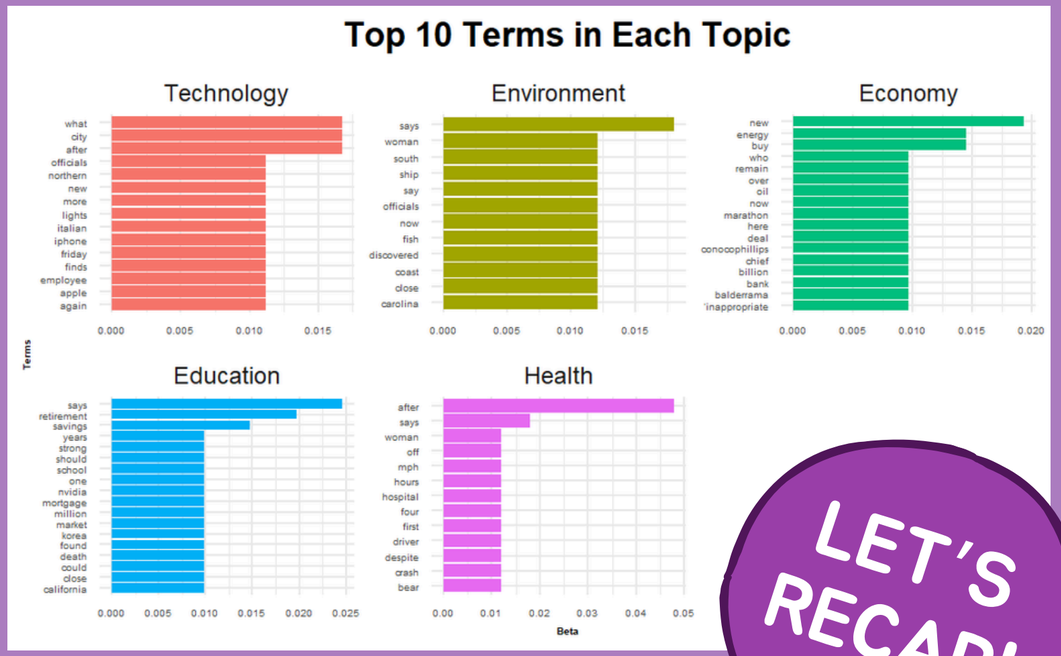
WordCloud

This Wordcloud provides a quick and informative visual overview of the main keywords and topics frequently covered in a document collection. By focusing on the biggest words, such as Swallows, Python, and Indonesia, we can understand the priority and direction of the information contained in the documents. This analysis is particularly useful for identifying key themes and assisting in decision-making regarding the focus of further research or analysis.

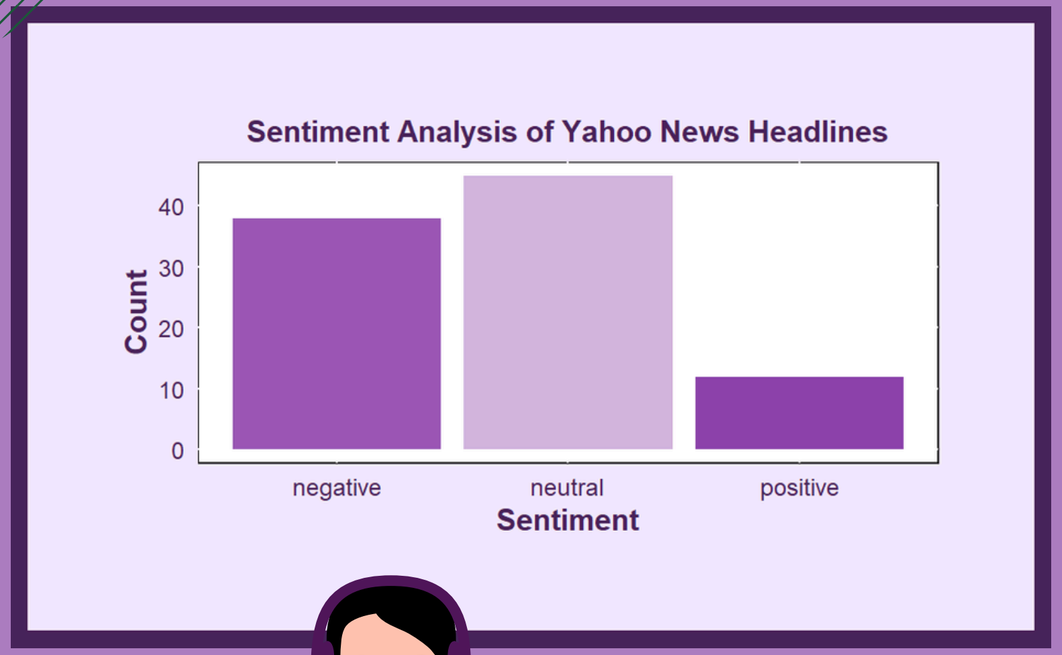
Latent Dirichlet Allocation

Analysis using the Latent Dirichlet Allocation (LDA) model on scraped data from Yahoo.com news reveals five main topics of interest and variety. The first topic, Technology, highlights the latest news from major technology companies like Apple, focusing on new products and related events. The second topic, Environment, covers issues of the natural environment and marine life, with news about coastal areas and discoveries at sea. The third topic, Economy, discusses global economic conditions, major business transactions, and developments in the energy industry, including oil and gas.

The fourth topic, Education, combines personal financial planning with the education sector, highlighting the importance of retirement savings and investments. Finally, the fifth topic, Health, focuses on health and safety, covering news about accidents, emergencies, and medical issues. The analysis provides a comprehensive overview of the key issues raised in the news, helping readers understand the key trends and topics being discussed.



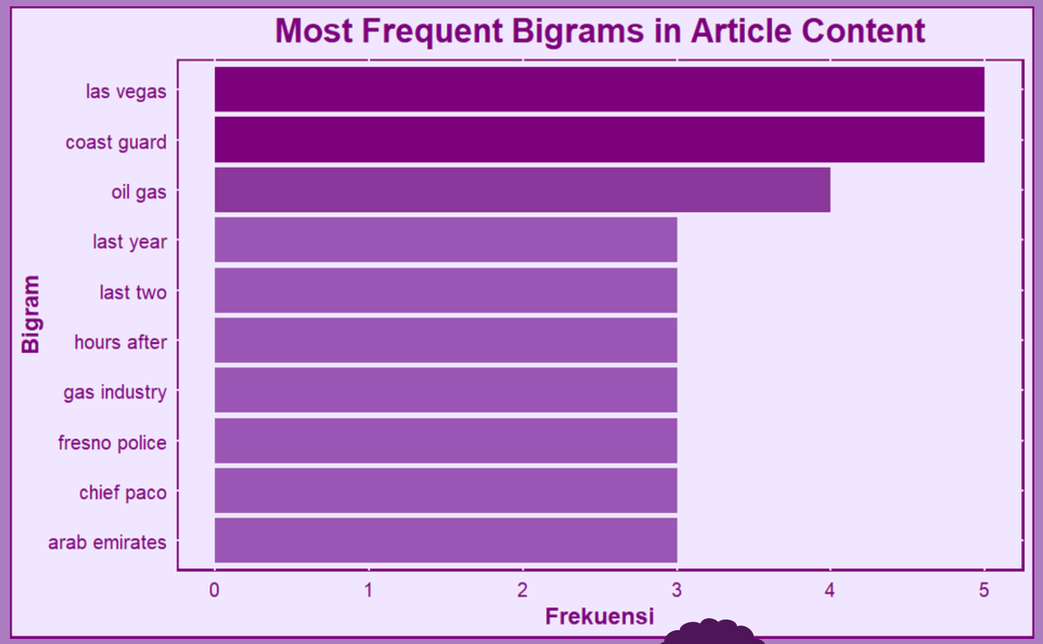
Sentiment Analysis of Yahoo!news Headlines



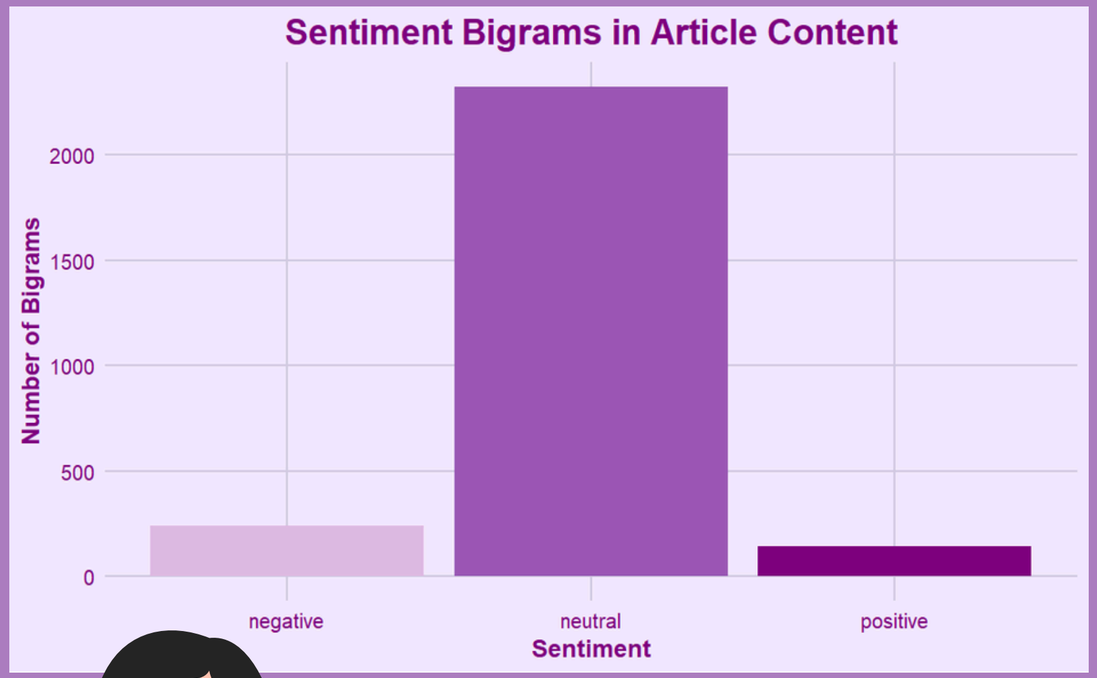
Sentiment analysis of Yahoo's headlines revealed that most news stories had neutral sentiment, followed by negative sentiment, and only a small percentage had positive sentiment. This suggests that the news presented tends to be more informative and objective, with a significant focus on issues with a negative impact, such as criminal events and social problems. Conversely, the rarity of positive news indicates that good news or news that inspires optimism is under-reported. This pattern illustrates how the media shapes public perceptions, emphasizing neutral information and negative news more than positive news.

Most Frequent Bigrams in Article Content

In analyzing the Most Frequent Bigrams in Article Content, some bigrams emerged that often dominated discussions, reflecting the hot topics that made headlines in various media. The bigram "Las Vegas" had the highest frequency, indicating that Las Vegas-related topics are frequently discussed. "Coast Guard" and "oil gas" also frequently appeared, focusing on news related to the Coast Guard and the oil and gas industry. Bigrams such as "last year" and "last two" are often used in the context of time, while "hours after" indicates the chronology of events. The bigram "Fresno police" indicates many articles related to the Fresno police, and "arab emirates" highlights topics regarding the Arab Emirates. This analysis helped identify key topics and themes in the articles, providing insight into the focus of the news stories.



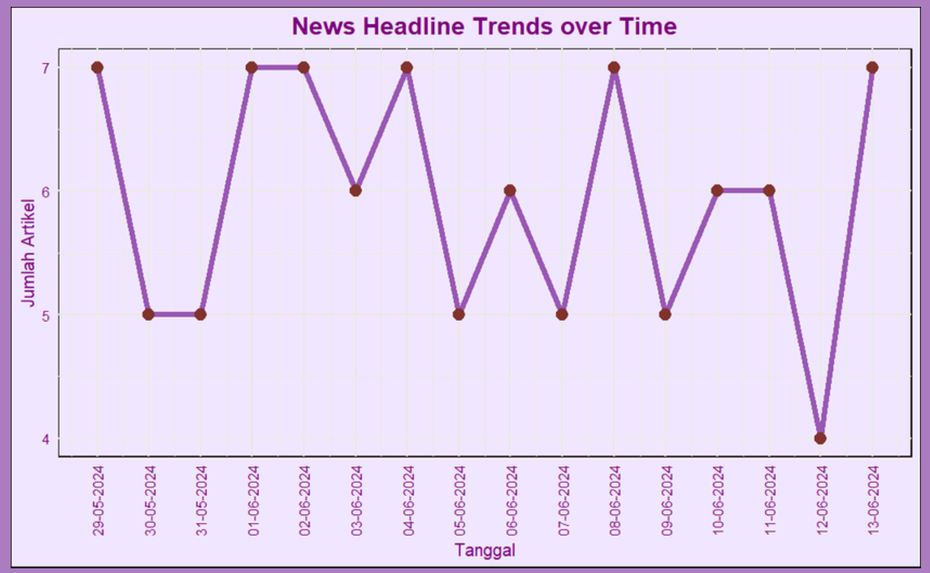
Sentiment Bigrams in Article Content



From this analysis, it can be concluded that article content tends to be more neutral, with less emphasis on positive and negative sentiments. This reflects an attempt to maintain balance in news reporting, although negative topics are still more prominent than positive ones. This phenomenon could be due to various factors, including the media's tendency to highlight shocking or unsettling events, which naturally attract greater attention from readers.

Additionally, the high number of neutral bigrams suggests that many articles focus on conveying information and facts without including opinions or emotions, which is important in maintaining credibility and reader trust. This visualization provides important insights into how language and phrases are used in articles, helping understand the sentiment bias in the media.

News Headline Trends over Time



Yahoo News' daily news distribution trend graph from May 29, 2024, to June 13, 2024, shows interesting fluctuations in the number of articles published daily. With peaks of up to 7 articles published on certain days, such as May 29 and June 13, and lows of 4 articles on other days, this graph reflects the media's responsiveness to significant events during these times.

These fluctuations indicate the high dynamics in the news world, where the intensity of news changes according to important events. The consistency of the number of articles on some days shows the editorial team's efforts to maintain a steady flow of information, while the peaks and valleys on the graph illustrate the adaptation to big and small news stories that emerge.

Conclusion

LIVE NEWS

Overall, it can be concluded that article content tends to be neutral, focusing more on balanced and objective information. Negative topics are still more prominent than positive ones, which may reflect the media's tendency to highlight shocking or unsettling events. This analysis provides deep insights into the use of language and phrases in news articles, helps understand sentiment bias in the media, and provides guidance for further research into how the media influences public perception.

BREAKING NEWS

THANK YOU!!

Meisyatul Ilma - G1501231073