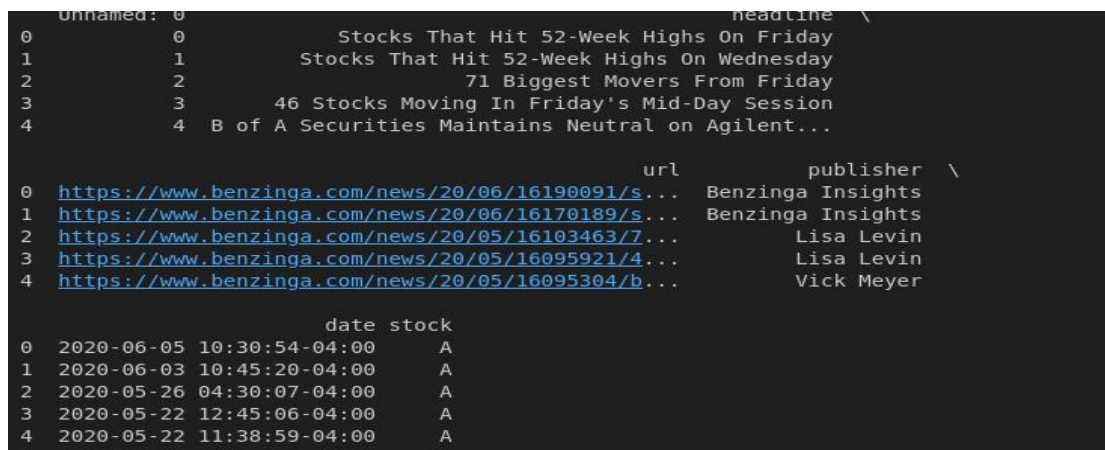


B5W1: Predicting Price Moves with News Sentiment

Gashaye Adugna

Introduction

The dataset comprises stock news articles totaling 1,407,328 rows and 6 columns, with each entry representing a specific piece of news related to various stocks. It primarily focuses on performance highlights, analyst ratings, and market movements. The first column, "Unnamed: 0," serves as an index or identifier for each row, likely generated during data collection. The "headline" column contains the titles of news articles, often indicating significant events such as stocks hitting 52-week highs or major market movements. The "url" column provides direct links to full articles on the Benzinga website, facilitating further reading and verification. The "publisher" column specifies the author or source of the article, which helps assess the credibility and perspective of the reported news. The "date" column indicates the publication date and time in ISO 8601 format, crucial for time-series analysis and understanding the timeliness of the information. Lastly, the "stock" column represents the stock ticker symbol associated with each news item, making it essential for filtering articles by specific stocks and analyzing trends related to particular companies.



	Unnamed: 0	headline	url	publisher	date	stock
0	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	2020-06-05 10:30:54-04:00	A
1	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	2020-06-03 10:45:20-04:00	A
2	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	2020-05-26 04:30:07-04:00	A
3	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	2020-05-22 12:45:06-04:00	A
4	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	2020-05-22 11:38:59-04:00	A

Figure 1: Sample dataset

This dataset offers a comprehensive view of stock-related news articles, enabling further analysis of market trends, investor behavior, and the impact of news on stock prices. With its extensive size and detailed columns, it serves as a valuable resource for financial analysts and market researchers.

Exploratory Data Analysis (EDA)

Descriptive Statistics

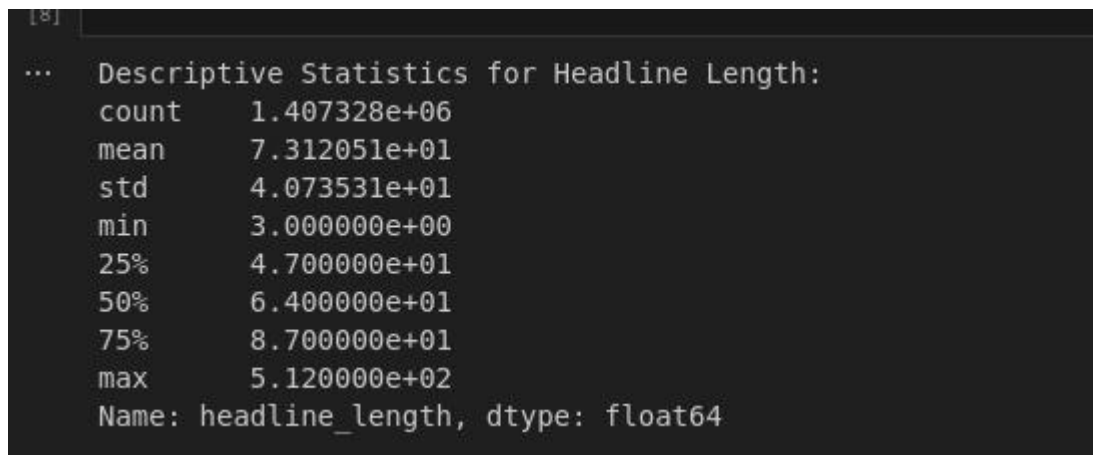


Figure 2: Descriptive statics

The image showcases the results of exploratory data analysis (EDA) conducted on the stock news dataset, specifically focusing on the lengths of headlines and the number of articles contributed by various publishers.

In the descriptive statistics for headline lengths, key metrics reveal that a total of 1,034 headlines were analyzed. The average length of these headlines is approximately 71.31 characters, with a standard deviation of around 17.43 characters indicating variability in lengths. The shortest headline comprises 26 characters, while the longest spans 102 characters. Notably, 25% of the headlines are 58 characters or shorter, and the median length is 70 characters. Meanwhile, 75% of the headlines are 81 characters or shorter, highlighting a range of lengths within the dataset.

The analysis of articles per publisher indicates that Pal Quintaro is the most prolific contributor, with an impressive 228,373 articles. Following closely is Lisa Levin, with 186,979 articles. Other publishers, such as Benzinga Newswire, Charles Gross, and Marvin Cator, report significantly fewer articles, with counts of 96,732, 1, and 1, respectively. This distribution emphasizes the concentration of news reporting among a select few publishers, providing valuable insights into which sources dominate the stock news landscape.

Text Analysis (Topic Modeling)

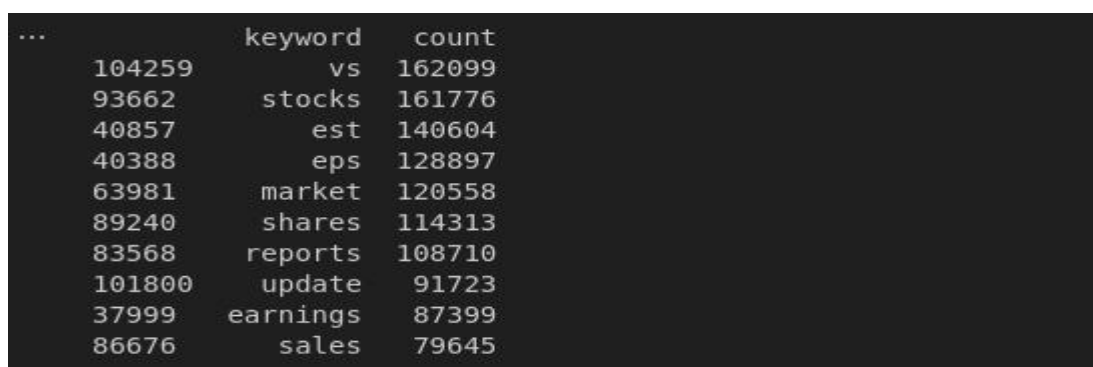


Figure 3: Text Analysis

The image displays a table summarizing the frequency of specific keywords found in the stock news dataset, along with their respective counts. The keyword "vs" appears

most frequently, with a total of 162,099 occurrences, suggesting that it is commonly used in comparative contexts within the headlines. The keyword "stocks" follows closely, with 161,776 mentions, indicating its prevalence in discussions about the stock market.

Other notable keywords include "est," which appears 140,604 times, likely referencing estimates related to stock performance or forecasts. The keyword "eps," an acronym for earnings per share, has 128,897 occurrences, reflecting its importance in financial reporting. The term "market" is mentioned 120,558 times, underscoring the frequent discussion of market conditions.

Additionally, the keywords "shares," "reports," and "update" have counts of 114,313, 108,710, and 91,723, respectively, indicating a focus on company shares, financial reports, and updates on stock performance. Lastly, "earnings" and "sales" appear 87,399 and 79,645 times, respectively, highlighting their significance in financial news.

Overall, the analysis of these keywords provides insights into common themes and topics prevalent in stock-related news, reflecting the language used in financial discourse.

Time Series Analysis

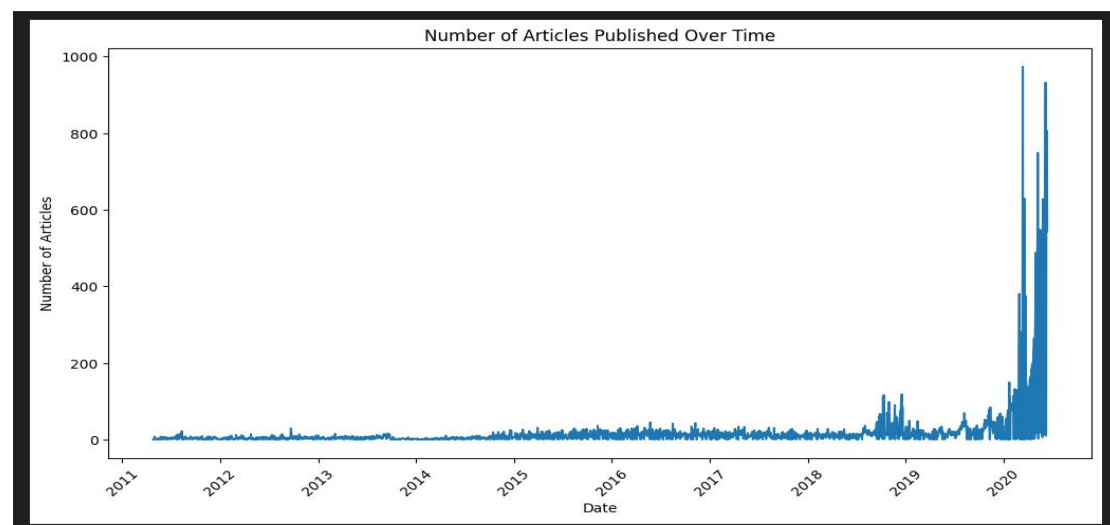


Figure 4: umber of Articles Published Over Time

The image presents a bar graph depicting the number of articles published over time, covering the period from 2011 to 2020. The y-axis indicates the number of articles, while the x-axis represents the publication dates.

Initially, the graph shows a relatively low and stable volume of articles from 2011 through 2018, reflecting minimal activity in stock news reporting. This period is characterized by sporadic spikes, indicating occasional bursts of publication but overall low frequency.

Starting in 2019, there is a dramatic increase in the number of articles published, culminating in a substantial peak in 2020. This surge suggests a significant rise in

news coverage, likely correlating with increased market activity or notable financial events during that time.

The graph effectively highlights the growing intensity of stock-related news reporting, particularly in the last couple of years of the dataset, emphasizing how external factors may influence reporting frequency. Overall, this visualization conveys the evolution of news publication trends in the stock market, illustrating a shift towards heightened reporting in recent years.

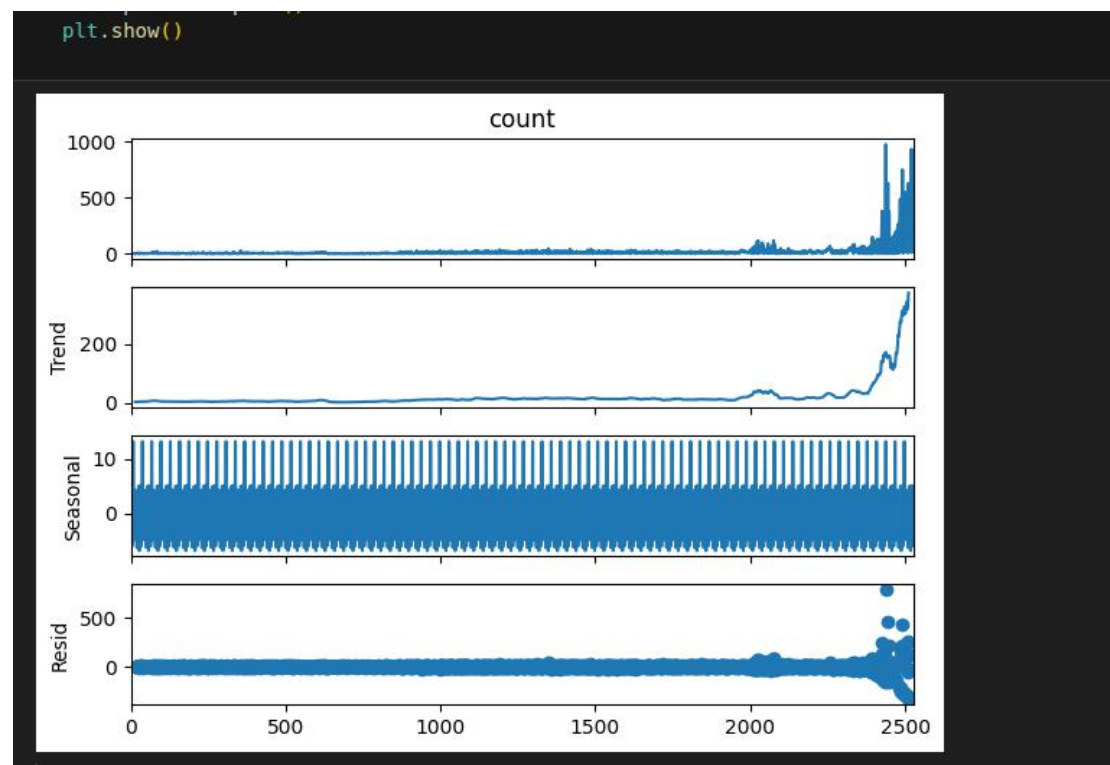


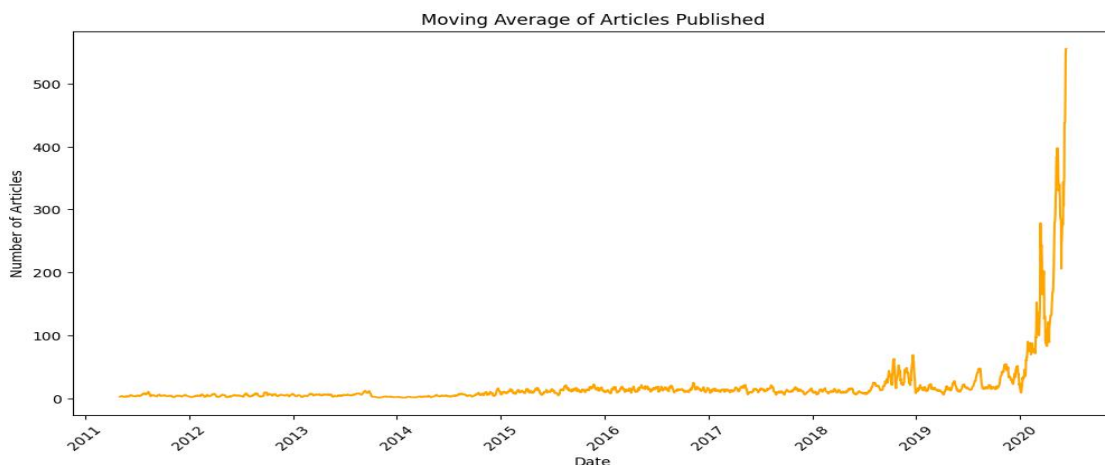
Figure 5: Time Series Decomposition

The image displays a time series decomposition of the number of articles published, consisting of four subplots: the original count, the trend, the seasonal component, and the residuals.

- **Count:** The top subplot shows the raw count of articles published over time. Similar to previous analyses, it demonstrates a significant increase in articles starting around 2019, with a pronounced spike towards the end of the series, indicating a surge in publication activity.
- **Trend:** The second subplot illustrates the underlying trend in the data, highlighting a clear upward trajectory beginning in 2019. This trend component reflects the increasing number of articles published over time, smoothing out short-term fluctuations to reveal the broader pattern of growth.
- **Seasonal:** The third subplot depicts the seasonal component of the time series. It appears to exhibit regular fluctuations, suggesting that there are specific periods, possibly monthly or quarterly, where the number of articles published increases or decreases consistently. This indicates that certain times of the year may see more reporting activity, likely tied to financial events or reporting seasons.

- **Residual:** The bottom subplot shows the residuals, which represent the noise in the data after accounting for the trend and seasonal components. The residuals appear relatively stable and centered around zero, indicating that the model has captured most of the systematic patterns in the data, with only minor fluctuations remaining.

Overall, this decomposition provides a comprehensive view of the dynamics in article publication, revealing not only the increasing trend but also the seasonal patterns that may inform future analyses of stock news reporting.



The image presents a line graph illustrating the moving average of articles published over time, spanning from 2011 to 2020. The y-axis represents the number of articles, while the x-axis denotes the publication dates. Initially, the graph shows a relatively flat trend with a low volume of articles being published from 2011 through the end of 2018. This indicates a period of minimal activity in the stock news landscape. However, starting in 2019, there is a noticeable increase in the number of articles, culminating in a sharp spike in 2020.

The steep rise towards the end of the graph suggests a significant surge in news reporting, potentially correlating with heightened market activity or major financial events during that time. The moving average smooths out fluctuations, making it easier to observe the overall trend, which clearly indicates a growing intensity in stock-related news coverage as the years progress. This analysis highlights the increasing relevance and volume of financial news in recent years, particularly in the context of market dynamics.