Exploratory Data Analysis (EDA) Report

Dataset: Financial News and Stock Price Integration Dataset (FNSPID)

Author: Chalie Lijalem

Date: 6/01/2025

Dataset Description

The Financial News and Stock Price Integration Dataset (FNSPID) is a multi-source dataset designed to assist in stock market predictions by integrating **qualitative news headlines** with **quantitative stock ticker information**. It contains the following fields:

- **headline**: Title of the news article.
- **url**: Link to the full article.
- **publisher**: Name of the news publisher or author.
- **date**: Publication date and time (UTC-4 timezone).
- **stock**: Stock ticker symbol (e.g., AAPL for Apple).

	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

Environment and Version Control Setup

To ensure proper development practices, the following were established:

- A GitHub repository was created to host all code related to this project.
- A new branch task-1 was created to track this analysis.
- The folder structure followed best practices (.vscode, src, notebooks, scripts, tests, etc.).
- Frequent commits were made with descriptive messages.
- A GitHub Actions CI pipeline (.github/workflows/unittests.yml) was added for automated testing.
- Python environment managed with requirements.txt.

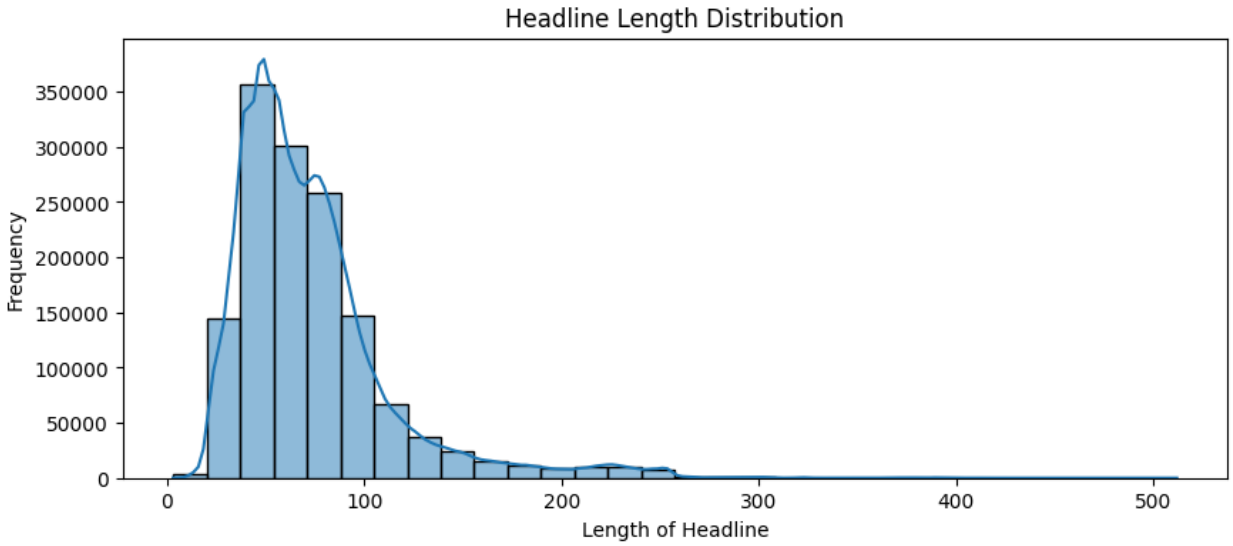
Descriptive Statistics

Headline Length

Basic statistics were computed for the length of article headlines. A histogram revealed a right-skewed distribution, indicating that while most headlines are concise, a few are significantly longer.

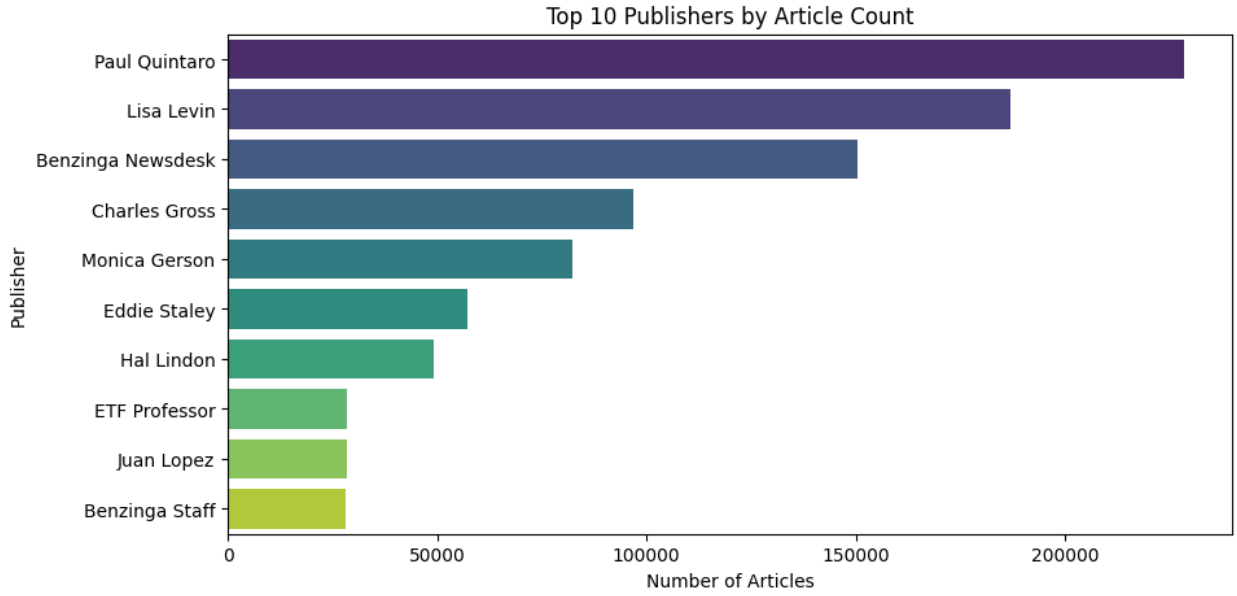
```
count    1.407328e+06
mean     7.312051e+01
std      4.073531e+01
min      3.000000e+00
25%      4.700000e+01
50%      6.400000e+01
75%      8.700000e+01
max      5.120000e+02
Name: headline_length, dtype: float64
```

This insight is crucial for downstream tasks like modeling or keyword extraction, where length may influence model performance.



Publisher Activity

An analysis of publishers revealed the top 10 contributors by article count. The most prolific publishers included **Paul Quintaro, Lisa Levin and Benzinga**, suggesting potential bias or dominance in the dataset. Understanding this helps identify authoritative or frequently quoted sources.



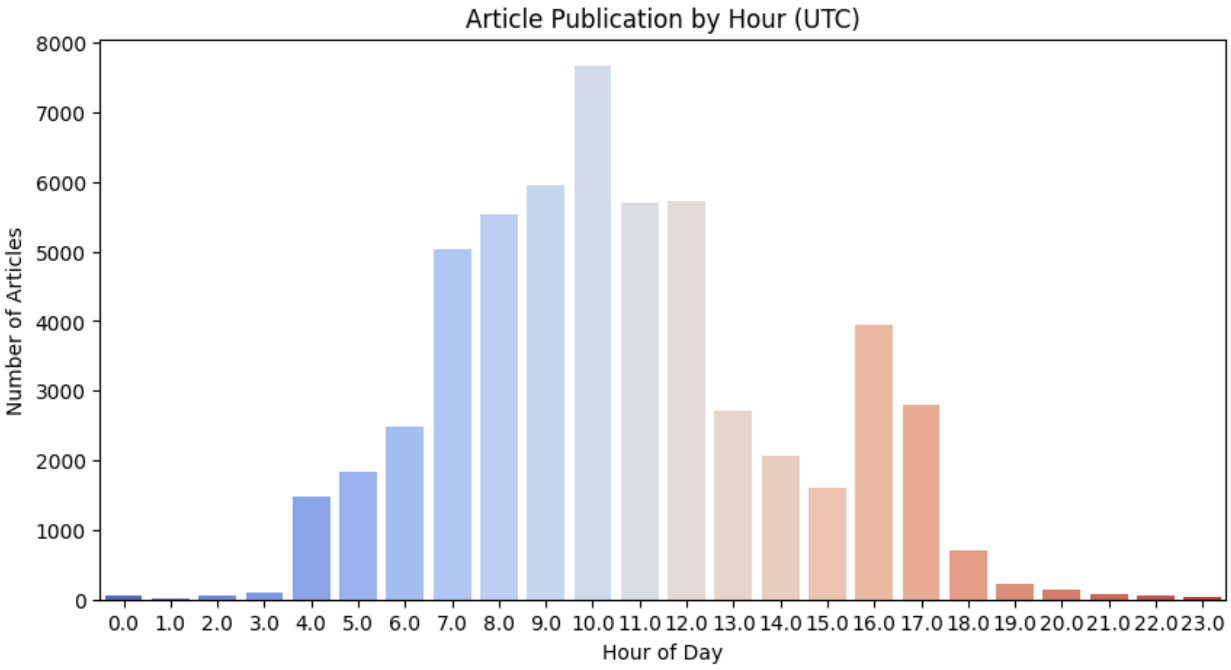
Temporal and Time Series Analysis

Articles Over Time

By aggregating publication counts over time, we observed noticeable spikes in publication volume. These spikes often correlate with high-impact market events such as earnings reports, regulatory changes, or macroeconomic announcements.

Hourly Publication Trends

An analysis of the publication hour (in UTC) showed that articles were most frequently released, which aligns with typical trading hours and news cycles in the U.S. market.



These trends are vital for building time-sensitive trading algorithms or alert systems.

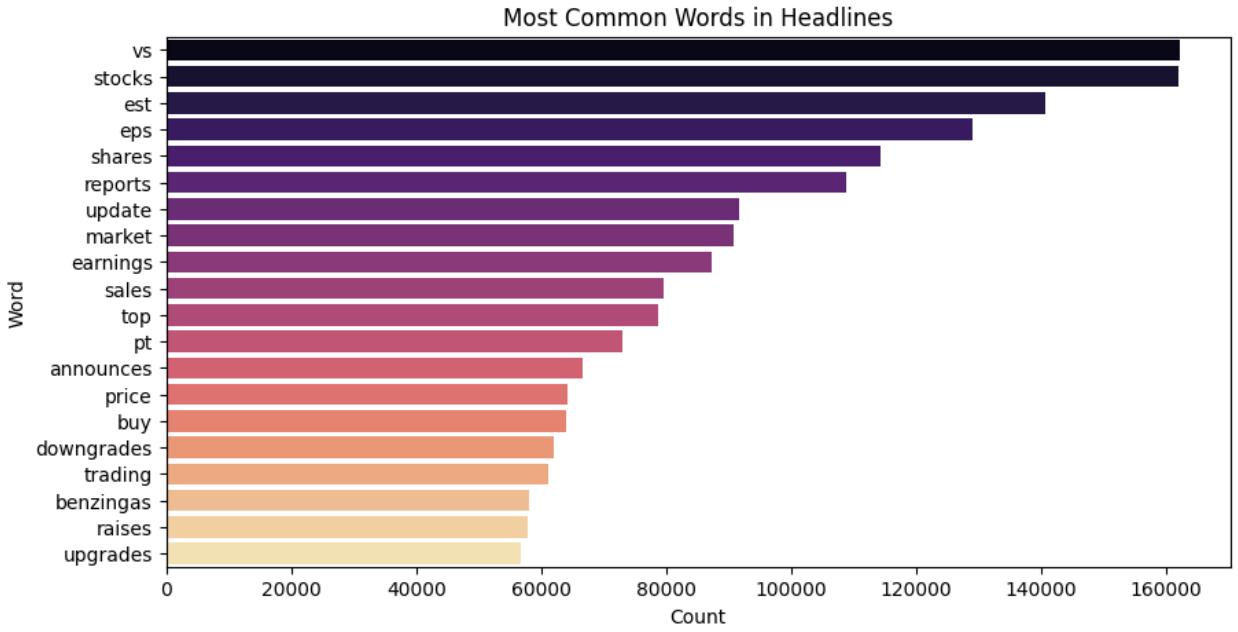
Text Analysis

Keyword Analysis

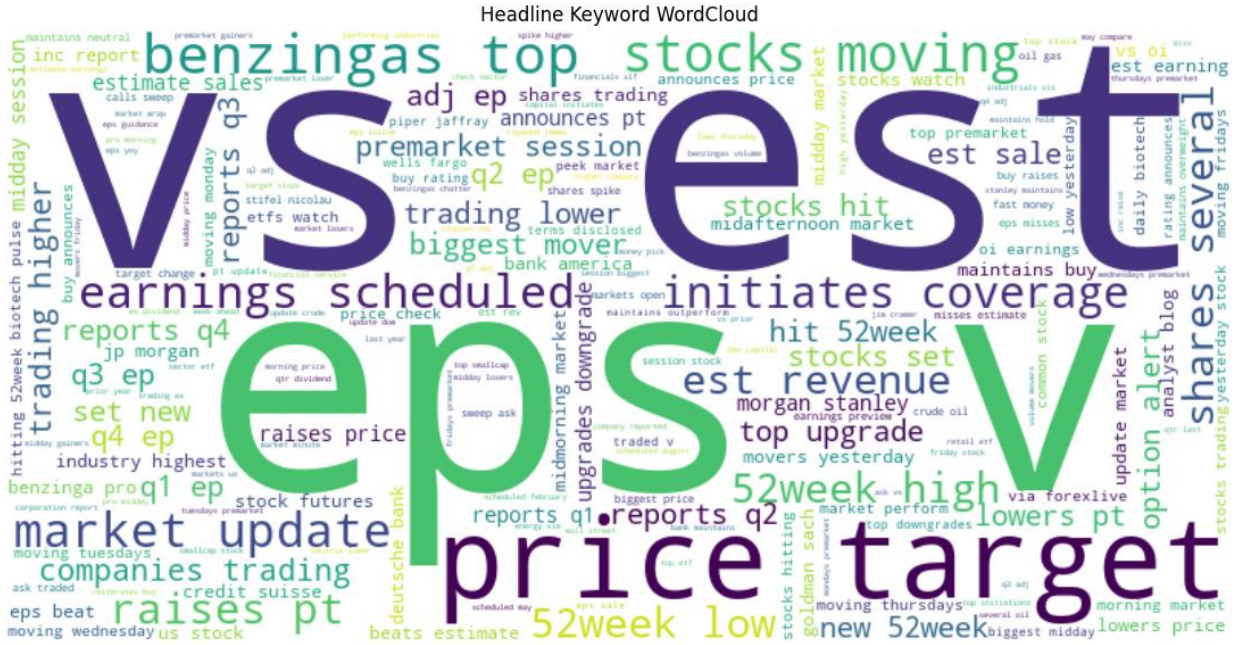
Text preprocessing was performed on the headlines using basic NLP techniques. The most frequent keywords included:

- *price*, *target*, *stock*, *shares*, *upgraded*, *downgraded*

These keywords reveal the dominant themes around price changes, recommendations, and market movements.



A **WordCloud** was generated to visually illustrate prominent terms. This helps traders and analysts quickly grasp headline content trends.



Topic Modeling (LDA)

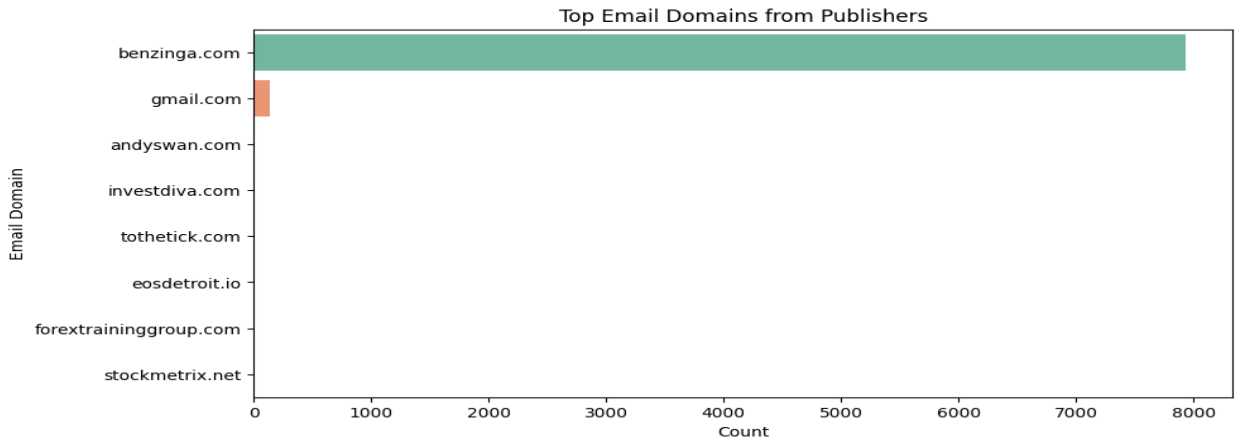
To identify underlying themes, **Latent Dirichlet Allocation (LDA)** was applied to the tokenized headlines. The model generated 5 key topics, each consisting of high-frequency keywords. For example:

- **Topic 1**: [stock, upgrade, buy, analyst, price] → Analyst recommendations
- **Topic 2**: [quarter, earnings, beat, revenue, estimate] → Earnings reports

This unsupervised topic modeling provides an abstract representation of the types of financial news covered and can be used for clustering or categorization tasks.

Publisher Domain Analysis

In cases where email addresses were used as publisher identifiers, the domain names were extracted. A handful of domains were disproportionately represented, indicating a few organizations contributed a large portion of the content. This domain analysis could help in future filtering or reliability assessments of publishers.



Conclusion

The EDA process has successfully revealed critical insights about the structure and characteristics of the FNSPID dataset:

- **Temporal trends** in article frequency and timing
- **Textual patterns** in headlines including dominant keywords and themes
- **Publisher dynamics**, both in terms of quantity and type
- **Readiness** of the data for downstream machine learning or financial prediction tasks

These insights not only aid in data preprocessing and feature engineering but also inform better decision-making in financial forecasting systems.