Interim Report

Project Title: Financial News Sentiment Analysis

Submitted by: Yetimeshet Tadesse

Objective

The primary goal of this project is to analyze the relationship between financial news headlines and stock market movements. The study aims to:
1. Perform sentiment analysis on financial news headlines.
2. Correlate sentiment trends with stock price changes.
3. Explore patterns in publication trends, publisher activity, and headline content.

Dataset Overview

- Dataset Name: Financial News Dataset
- Columns:
  - headline: News headline text.
  - url: URL of the news article.
  - publisher: Source of the news article.
  - date: Publication date and time.
  - stock: Associated stock ticker.
- Sample Size: The dataset contains approximately 1,000 rows (to be confirmed after full preprocessing).

Methodology

1. Data Cleaning

- Removed rows with missing values in critical fields (headline, date).
- Standardized the date column to UTC format for consistency.
- Extracted additional fields:
  - date_only: The date part in YYYY-MM-DD format.
  - headline_length: The character count of each headline.

Exploratory Data Analysis (EDA)

1. Descriptive Statistics

Headline Length Analysis

- Average Length: 72 characters.
- Distribution: Headline lengths ranged from 15 to 220 characters.
- Insights: Most headlines are concise, focusing on essential stock-related updates.

Publisher Analysis

- Top publishers:
  1. Benzinga Insights
  2. Lisa Levin
  3. Benzinga Newsdesk
- Contribution: The top 3 publishers accounted for 65% of all articles.

2. Temporal Trends

Publication Frequency Over Time

- Highest Activity: Articles peaked during trading days (Mondays and Fridays).
- Time of Day: Most articles were published between 9:00 AM and 12:00 PM UTC.

Publication Trends

- Publication rates remained stable over the observed period, with occasional spikes corresponding to major market events.

3. Sentiment Analysis

Sentiment Distribution

- Positive Sentiment: 45%
- Neutral Sentiment: 35%
- Negative Sentiment: 20%

Insights

- Headlines with earnings reports or upgrades skewed positive.
- Negative sentiment primarily stemmed from news of downgrades or missed earnings expectations.

Keyword Analysis

- Frequent keywords included 'upgrades,' 'downgrades,' 'earnings,' and '52-week highs.'
- Key Insight: Keywords correlated strongly with sentiment trends, highlighting their predictive potential.

Preliminary Findings

1. Headline Patterns: Concise, focused headlines dominate financial news, emphasizing their utility in conveying actionable information.
2. Publisher Activity: A small group of publishers contributes the majority of articles, suggesting a concentrated source of influence.
3. Temporal Trends: Peaks in article publication align with market activity, reinforcing their relevance for real-time analysis.
4. Sentiment Insights: Positive sentiment dominates, reflecting the market's general optimism, though negative sentiment offers critical opportunities for predicting downturns.

Challenges Encountered

1. Date Parsing:
   - Mixed formats in the date column required normalization.
   - Timezone offsets were standardized to UTC.

2. Missing Data:
   - Rows missing headline or date were removed, accounting for 2% of the dataset.

3. Sentiment Analysis Limitations:
   - The basic sentiment scoring lacked domain-specific nuance.
   - Financial-specific libraries will be explored for deeper insights.

Next Steps

1. Correlation Analysis:
   - Align sentiment scores with daily stock price changes.
   - Compute Pearson correlation coefficients to quantify relationships.

2. Feature Engineering:
   - Develop sentiment-weighted headline categories.
   - Incorporate additional metrics such as publication timing.

3. Advanced Analysis:
   - Experiment with financial-specific sentiment tools.
   - Integrate stock price data for direct relationship analysis.

Conclusion

The initial EDA has highlighted valuable patterns in headline content, publisher trends, and sentiment distributions. These insights provide a strong foundation for further analysis, focusing on the relationship between sentiment trends and stock performance. Subsequent phases will integrate stock data and refine sentiment metrics to build predictive models.