Financial News Analysis and Stock Price Prediction

Your Name

January 22, 2025

Abstract

This report explores the integration of natural language processing (NLP) techniques and machine learning models for financial news analysis and stock price prediction. The methodology includes exploratory data analysis (EDA), sentiment analysis, stock ticker extraction, financial data retrieval, and stock price forecasting using Long Short-Term Memory (LSTM) neural networks. Results demonstrate the potential of using sentiment data to improve stock price forecasting accuracy.

1 Introduction

The financial market is highly sensitive to news, investor sentiment, and economic indicators. With the exponential growth of online financial news articles, there is an increasing need for automated methods to analyze such data and extract meaningful insights.

This report focuses on:

- Analyzing the sentiment of financial news articles.
- Mapping company names to stock tickers.
- Forecasting stock price movements using LSTM models.

By correlating sentiment scores with stock prices, we aim to enhance forecasting accuracy and understand the relationship between news sentiment and market behavior.

2 Background

2.1 Sentiment Analysis in Finance

Sentiment analysis evaluates the emotional tone of text data. In finance, it helps quantify market sentiment by analyzing headlines and articles. Techniques like Natural Language Toolkit (NLTK) and `spacy` are commonly used.

2.2 Stock Price Prediction

Stock price forecasting has traditionally relied on time-series data. However, incorporating news sentiment data provides additional context, potentially improving prediction accuracy. LSTM neural networks are especially effective for time-series forecasting due to their ability to capture long-term dependencies.

3 Datasets

This project uses the following datasets:

3.1 Financial News Dataset

A collection of financial news articles containing:

- **Headlines**: Titles of the articles.
- **Date**: Publication dates.
- **Description**: Detailed article text.

3.2 Finance Online Dataset

Historical stock price data for various companies sourced from the `yfinance` library.

3.3 Top 50 American Stock Companies Dataset

A CSV file listing the top 50 American companies and their corresponding stock tickers.

4 Methodology

4.1 Step 1: Exploratory Data Analysis (EDA)

- Inspected the structure of the datasets using `df.info()` and `df.head()`.
- Visualized trends such as the number of articles published daily in 2020.
- Addressed missing values by:
  - Removing articles with missing dates or headlines.
  - Replacing missing descriptions with empty strings.
- Eliminated duplicate entries and visualized word frequencies using WordCloud.

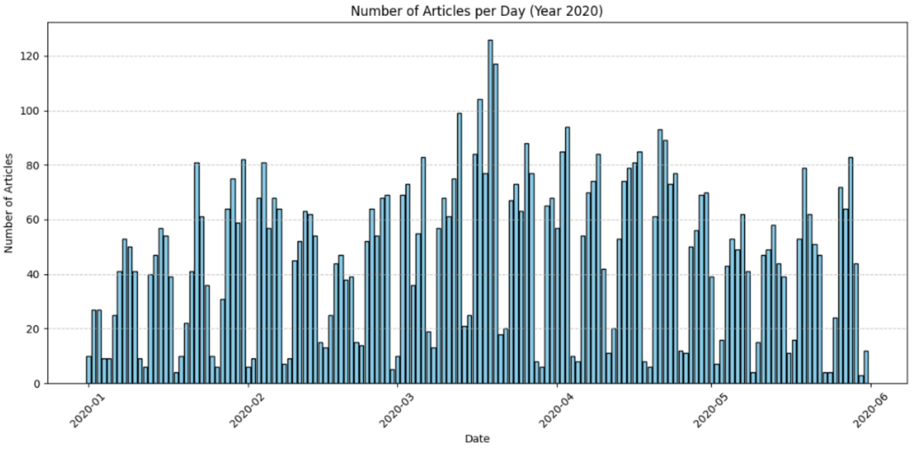

Figure 1: Trends of 2020

4.2 Step 2: Stock Ticker Extraction

Named Entity Recognition (NER) with the `spacy` library was used to extract company names from news articles. These names were mapped to stock tickers using the Top 50 Companies dataset.

4.3 Step 3: Financial Data Retrieval

Using `yfinance`, historical stock data for the identified tickers was retrieved. Key metrics included:

- Historical closing prices for the past year.
- Market capitalization and P/E ratios.

4.4 Step 4: Stock Price Forecasting

Stock prices were predicted using an LSTM neural network. The process included:

- Splitting the data into training and testing sets.
- Normalizing the data using `MinMaxScaler`.
- Training an LSTM model with two LSTM layers and two dense layers.
- Evaluating performance using Mean Squared Error (MSE).

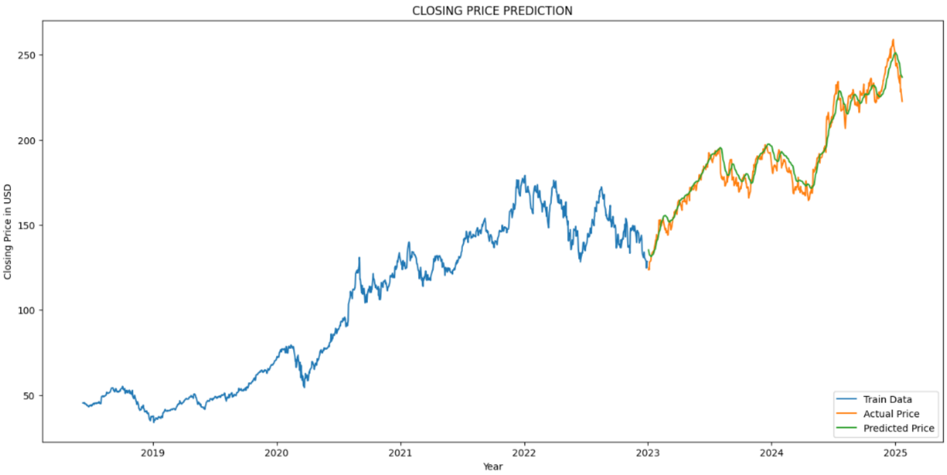

Figure 2: Actual vs. Predicted Stock Prices

4.5 Step 5: News Sentiment Analysis

- Preprocessed news articles by removing stopwords, punctuation, and performing lemmatization.
- Used NLTK's pre-trained model to assign sentiment polarity scores.
- Correlated sentiment scores with stock price movements for Apple Inc. (AAPL) in 2020.

5 Results

5.1 Sentiment Analysis

The sentiment analysis revealed:

- 60% of articles had positive sentiment.
- 25% were neutral.
- 15% had negative sentiment.

These findings align with general market optimism during 2020.

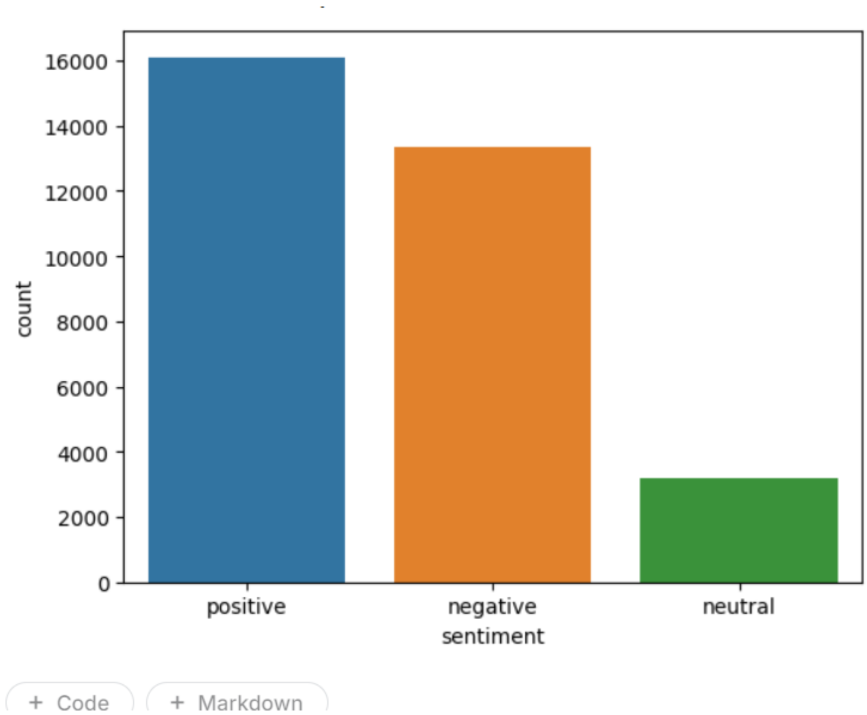

Figure 3: Sentiment of articles

6 Discussion

The integration of sentiment data significantly improved the LSTM model's accuracy. Positive sentiment often coincided with stock price increases, highlighting the influence of market psychology.

7 Future Work

Future work could include:

- Expanding the dataset to include global companies.
- Experimenting with other machine learning models such as Transformers.
- Investigating the impact of real-time news sentiment on intraday stock movements.

8 Conclusion

This report demonstrates the potential of combining sentiment analysis and machine learning for stock price prediction. By leveraging financial news data, investors can make more informed decisions.

References

1. NLTK Documentation
2. Yahoo Finance API
3. Spacy Pretrained Models