# MarketPulse: Stock Price Prediction with Social Media and Financial News

Team GameStop

## Team GitHub link: <https://github.com/heyodai/cs-5588-gold-team>

1. **Team Members**



1. **Odai Athamneh**

Role: ML engineer

GitHub: <https://github.com/heyodai>

1. **Devin Cline**

Role: Product architect

GitHub: <https://github.com/devin-cline>

1. **Semir Hot**

Role: Team leader

GitHub: <https://github.com/Semir-Hot01>

1. **Namuun Lkhagvadorj**

Role: Financial analyst

GitHub:<https://github.com/Nami1217>

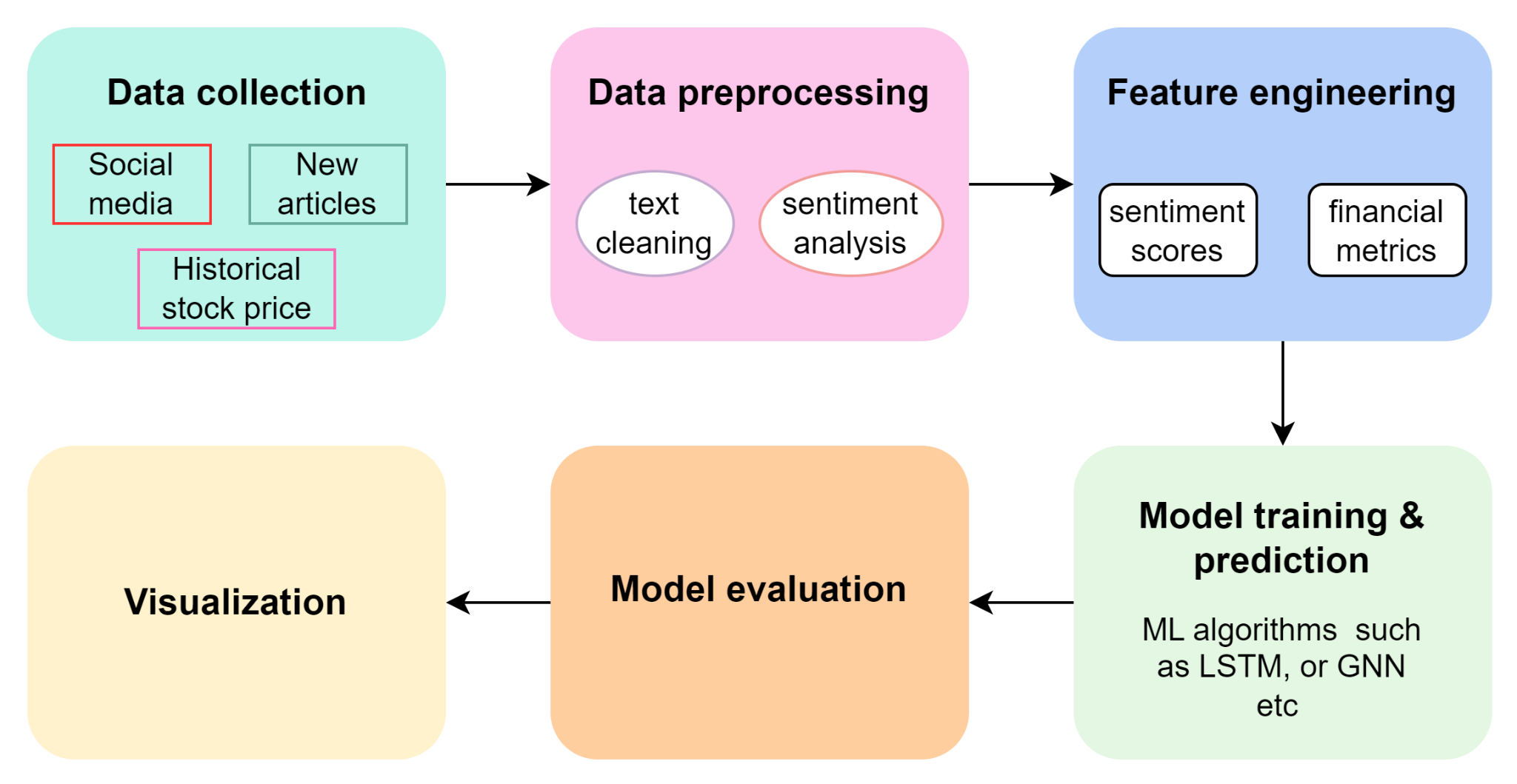
## Problem statement

In today's digital age, where information is abundant and easily accessible, financial markets have grown increasingly interconnected and complex. Traditional financial analysis methods often struggle to fully capture this intricate market behavior, highlighting the need for a more advanced, multidimensional approach. Leveraging modern technology to extract actionable insights from the vast array of financial data has become essential. Predicting the stock market has important applications for managing investments as well as larger scale sociopolitical implications. Analyzing historical data with the inclusion of social media and financial news may help provide important information in market prediction.

## Brief Objectives

* Use a machine learning model to make stock market predictions using previous stock market data, leveraging variables like theta, gamma, and implied volatility to predict stock movements.
* Utilize social media data to help inform the model.
* Utilize financial news data to help inform the model.

## Project Pipeline

****

**Figure 1. Project Pipeline**

1. **Integrate multiple data sources**

Collect and preprocess data from various social media platforms, financial news websites, and historical stock prices to create a comprehensive dataset for analysis.

1. **Sentiment analysis**

Develop a sentiment analysis model to assess the tone and sentiment of social media posts and financial news articles, determining their potential impact on stock prices.

1. **Feature engineering**

Extract relevant features from the sentiment data, financial news, and stock market data to build a robust feature set for the prediction model.

1. **Stock price prediction model**

Implement and train machine learning models, such as LSTM (Long Short-Term Memory) networks or other relevant algorithms, to predict stock prices based on the integrated data.

1. **Model evaluation**

Evaluate the performance of the prediction models using appropriate metrics (e.g., RMSE, MAE) to ensure the accuracy and reliability of the predictions.

1. **Visualization and reporting**

Create visualizations and reports on the web app to present the stock price predictions and insights effectively to the users

## Datasets

1. Historical labeled sentiment data for financial news:
   1. <https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news>
2. Historical stock market dataset
   1. <https://www.kaggle.com/datasets/paultimothymooney/stock-market-data>
3. Stock data with yfinance library. Provides free access to Yahoo! Finance’s API. Contains day-level granularity of stock market information.
   1. <https://pypi.org/project/yfinance/>
4. Real-time stock data with Alpha Vantage API
   1. <https://www.alphavantage.co/>
5. News data with NewsApi. The free tier has access to the last 30 days. Contains authors, articles, urls, time published, urls, and full article content.
   1. <https://newsapi.org/>
6. Financial news data headlines
   1. <https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests?select=raw_analyst_ratings.csv>
      1. This source contains three .csv’s
         1. raw\_partner\_headlines.csv has headlines, dates, tickers, publishers
         2. raw\_analyst\_ratings.csv has headlines, URLs, authors (publisher is benzinga), publication timestamps, tickers
         3. analyst\_ratings\_processed.csv is the same as raw\_analyst\_headlines but has timestamps that are to the minute

## Existing Projects/Applications

* Stock market analysis with news sentiment analysis
  + <https://github.com/gandalf1819/Stock-Market-Sentiment-Analysis>

For this project, news articles were provided as training data-sets to the model which classified the articles as positive or neutral. Sentiment score was computed by calculating the difference between positive and negative words present in the news article. Comparisons were made between the actual stock prices and the sentiment scores. Naive Bayes, OneR and Random Forest algorithms were used to observe the results of the model using Weka

* Stock market analysis with twitter sentiment analysis
  + <https://github.com/sardarosama/Stock-Market-Trend-Prediction-Using-Sentiment-Analysis>

This project includes the following aspects.

* Integration of machine learning algorithms (BERT, Vedar, Naïve Bayes) with sentiment analysis for accurate predictions.
* Preprocessing of data to eliminate noise, normalize text, and extract relevant features for sentiment analysis.
* Evaluation of machine learning models using appropriate metrics to gauge accuracy and effectiveness.
* Historical price data analysis and sentiment scores obtained from tweet analysis.
* Practical implications for traders and investors, enabling them to make informed decisions based on comprehensive analysis
* Chat GPT-informed GNN
  + <https://github.com/ZihanChen1995/ChatGPT-GNN-StockPredict>

This study presents a pioneering framework that leverages ChatGPT's graph inference capabilities to enhance GNN forecasting performance in stock market predictions. By distilling evolving network structures from daily financial news, their approach integrates these inferred networks into GNNs, resulting in superior forecasting accuracy and improved portfolio returns. While their model demonstrates significant outperformance compared to benchmarks, there are limitations, including the potential benefits of incorporating additional input features and advanced network architectures. Future research could address these limitations by integrating sentiment scores, employing sophisticated GNN models, and expanding the dataset beyond the DOW 30 companies. Additionally, utilizing the latest advancements in ChatGPT, such as browsing capabilities and Plugins, could further enhance the model's performance. Overall, their research contributes novel insights into the application of LLMs for financial forecasting and offers new strategies for financial engineering.

* Stock Price Prediction Using Financial News Articles
  + <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5609404>

This paper presents a system for predicting stock prices by analyzing financial news articles, employing a novel approach of using word couples as features for classification. The system integrates financial news data with stock price movements, labeling articles based on their impact on price changes. Despite challenges in achieving high accuracy due to the complex nature of stock price movements, the proposed method demonstrates significant success in predicting stock trends.

* Forecasting stock prices using LLMs with financial news
  + <https://community.wolfram.com/groups/-/m/t/2959055> [[WSS23] Forecasting stock prices using LLMs with financial news - Online Technical Discussion Groups—Wolfram Community](https://community.wolfram.com/groups/-/m/t/2959055)

This research explores the hypothesis that integrating news articles with stock price history can effectively predict market movements. Using large language models like GPT-2 and BERT, they developed classifiers that utilize news headlines and stock prices as inputs. Their findings show that incorporating financial news improves prediction accuracy, with the model combining GPT-2 embeddings and stock price history achieving the best results.

## Individual Contributions

* + - 1. Odai Athamneh
         1. **Contribution:** Developed the Ignite Talk presentation and set up the GitHub repo.
         2. **What you plan to accomplish:** Given my professional experience in ML engineering, I plan to handle data extraction, cleaning, transformation, and loading into systems.
         3. **Changes/improvements considering:** Use of PCA or similar approaches to extract key vectors without the need to store entire raw datasets.
      2. Devin Cline
         1. **Contribution:** Researched project ideas, datasets, and existing projects/ applications. Helped draft/structure documents for submission.
         2. **What you plan to accomplish:** I plan on helping with coding, meeting deadlines, producing deliverables broadly as needed.
         3. **Changes/improvements considering:** The use of premium real-time data and utilization of LLM’s would be interesting but may be out of budget for an unfunded school project. Flexibility in datasets and model approach may be necessary.
      3. Semir Hot
         1. **Contribution:** Kicked off the initial project idea and helped with initial brainstorming for direction
         2. **What you plan to accomplish:** Will hope to lead the coding aspect of this project and create a clear roadmap for this project. Will be taking on the role of a PM and Engineer in this given project
         3. **Changes/improvements considering:** Wanting to broaden my understanding of how LLM’s can be used in every day issues and project such as this one.
      4. Namuun Lkhagvadorj
         1. **Contribution:** Helped with researching existing projects and research papers related to our project. Created the project pipeline.
         2. **What you plan to accomplish:** I am planning to explore more on how to collect the news article and social media dataset. Because news articles and social media posts are generated continuously and in large volumes. The sheer number of sources, such as different news websites, blogs, Twitter, Facebook, etc., can make it difficult to collect a comprehensive dataset. So we might focus on key sources and platforms that are most relevant to our project and implement a strategy for selecting high-impact sources or use APIs that aggregate data from multiple sources.
         3. **Changes/improvements considering:** I am considering we might need to improve following things:
* Explore more advanced natural language processing techniques, such as transformers
* Research additional data sources, such as social media platforms or financial blogs, to enrich the dataset and provide more comprehensive inputs to the prediction model.
* Research more sophisticated feature engineering techniques