CapX Assignment

Name: Dharmanshu Singh
Contact: 8303550486

Table Of Content

Stock Movement Analysis Based on Social Media Sentiment

Objective: Develop a machine learning model that predicts stock movements by scraping data from social media platforms like Twitter, Reddit, or Telegram. The model should extract insights from user-generated content, such as stock discussions, predictions, or sentiment analysis, and accurately forecast stock price trends.

1. Github link of Project
2. Streamlit App link
3. Youtube Demo Link

Introduction:

CapX is an advanced Python-based project that leverages machine learning and natural language processing to extract stock market insights from Reddit discussions. By combining web scraping, sentiment analysis, and predictive modeling, CapX offers a comprehensive approach to understanding stock market trends through social media analysis.

Methodology: Detailed Approach

The *Stock Movement Sentiment Predictor* project follows a structured pipeline to analyze social media data, extract valuable insights, and predict stock movements. This methodology ensures a balance between data preprocessing, feature engineering, and model development. Below is a comprehensive breakdown of each stage in the process.

---

1. Data Collection

The data for this project was sourced from **Reddit's financial discussions**, particularly posts and comments in subreddits like *r/WallStreetBets* and *r/StockMarket*.

- **Structure of Data**:
  - Each post contained attributes such as the title, body, timestamp, engagement metrics (upvotes, comments), and ticker symbols.
  - Relevant metadata like post popularity and user interactions were also collected.
- **Data Size**:
  - The dataset comprised thousands of posts spanning several months, ensuring sufficient coverage of market movements and discussions.
  - Each data point was labeled as "Up," "Down," or "Neutral" based on the corresponding stock's price movement on the following trading day.
- **APIs and Tools**:
  - Data was retrieved using Reddit's API, leveraging Python libraries like `PRAW` for efficient scraping and retrieval.

---

2. Data Preprocessing

Raw social media data is often noisy and unstructured, requiring robust preprocessing steps to make it suitable for analysis. Key preprocessing tasks included:

1. **Text Cleaning**:
    - Removed unwanted elements such as HTML tags, URLs, special characters, and emojis.
    - Example:
      Input: *"🚀 Apple stocks are going to the moon! Check this out: http://example.com"*
      Output: *"Apple stocks are going to the moon"*
2. **Case Normalization**:
    - Standardized all text to lowercase for consistent processing.
3. **Stopword Removal**:
    - Eliminated common words like *"the," "and," "is"* that do not contribute meaningfully to sentiment analysis.
4. **Lemmatization**:
    - Transformed words to their base forms (e.g., *"running"* → *"run"*) to reduce dimensionality and focus on semantics.
5. **Ticker Symbol Identification**:
    - Extracted stock ticker symbols using regular expressions, ensuring relevant posts were linked to the corresponding stock.
6. **Outlier Handling**:
    - Filtered posts with extremely low engagement (e.g., posts with 0 comments/upvotes) to focus on impactful discussions.

---

3. Sentiment Analysis

Sentiment analysis was conducted to quantify the emotions and opinions expressed in posts. This step was pivotal in understanding market sentiment.

1. **Polarity Scoring**:
    - Used the **VADER Sentiment Analyzer** to assign polarity scores (ranging from -1 to +1):
        - Positive sentiment (e.g., bullish posts predicting stock price hikes)
        - Negative sentiment (e.g., bearish posts warning of stock declines)
        - Neutral sentiment (e.g., general discussions or news summaries)
2. **Aggregation**:

    - For posts mentioning multiple stocks, sentiment scores were aggregated for each ticker symbol to avoid bias.
3. **Validation**:
    - Sample sentiment predictions were manually cross-checked to ensure accuracy.

---

4. Topic Modeling

Topic modeling was used to identify key themes in discussions, providing deeper insights into what drives stock sentiment.

1. **Latent Dirichlet Allocation (LDA)**:
    - Implemented LDA to extract hidden topics from post text.
    - Topics such as *"earnings reports," "market crashes," "mergers," and "regulatory news"* were identified.
2. **Visualization**:
    - Visualized topic clusters using tools like **pyLDAvis**, helping to interpret dominant themes in the dataset.

---

5. Feature Engineering

To enhance the predictive power of the model, various features were engineered from the raw data:

1. **Sentiment Polarity**:
    - Numerical representation of sentiment, directly derived from the VADER analyzer.
2. **Engagement Metrics**:
    - Features like upvotes, comments, and post length were included to capture the popularity and impact of posts.
3. **Temporal Features**:
    - Captured the day and time of post creation, as stock sentiment can vary across market hours and weekends.
4. **Textual Features**:
    - Title length, keyword density (e.g., frequency of words like "buy," "sell"), and readability scores were calculated.
5. **Ticker Frequency**:
    - Counted how often specific stocks were mentioned, with higher counts indicating stronger market attention.

---

6. Model Development

The core predictive model was a **Random Forest Classifier**, chosen for its ability to handle complex relationships and high-dimensional data effectively.

1. **Model Selection**:
    - Random Forests were selected after experimenting with several classifiers, including Logistic Regression, Support Vector Machines, and Gradient Boosting.
    - Random Forests consistently outperformed others in terms of accuracy and interpretability.
2. **Hyperparameter Tuning**:
    - Used **GridSearchCV** to optimize parameters like the number of trees, maximum depth, and minimum samples per split.
3. **Data Splitting**:
    - Split the dataset into training (80%) and testing (20%) subsets.
    - Ensured class balance using stratified sampling to avoid biased predictions.

---

7. Model Evaluation

The trained model was rigorously tested to assess its predictive capabilities:

1. **Confusion Matrix**:
    - Evaluated the model's ability to correctly classify stock movements as "Up," "Down," or "Neutral."
2. **Performance Metrics**:
    - Accuracy, Precision, Recall, and F1-Score were calculated to provide a holistic evaluation of the model's performance.
3. **Cross-Validation**:
    - Performed 5-fold cross-validation to ensure the model's robustness and reliability across different data subsets.

---

8. Deployment

Finally, the model was integrated into a user-friendly interface, allowing real-time predictions:

1. **Dashboard**:
    - Built an interactive dashboard using **Streamlit**, enabling users to upload social media datasets and view predictions.
2. **Scalability**:
    - Ensured the pipeline could handle larger datasets, incorporating additional data sources in the future (e.g., Twitter, StockTwits).

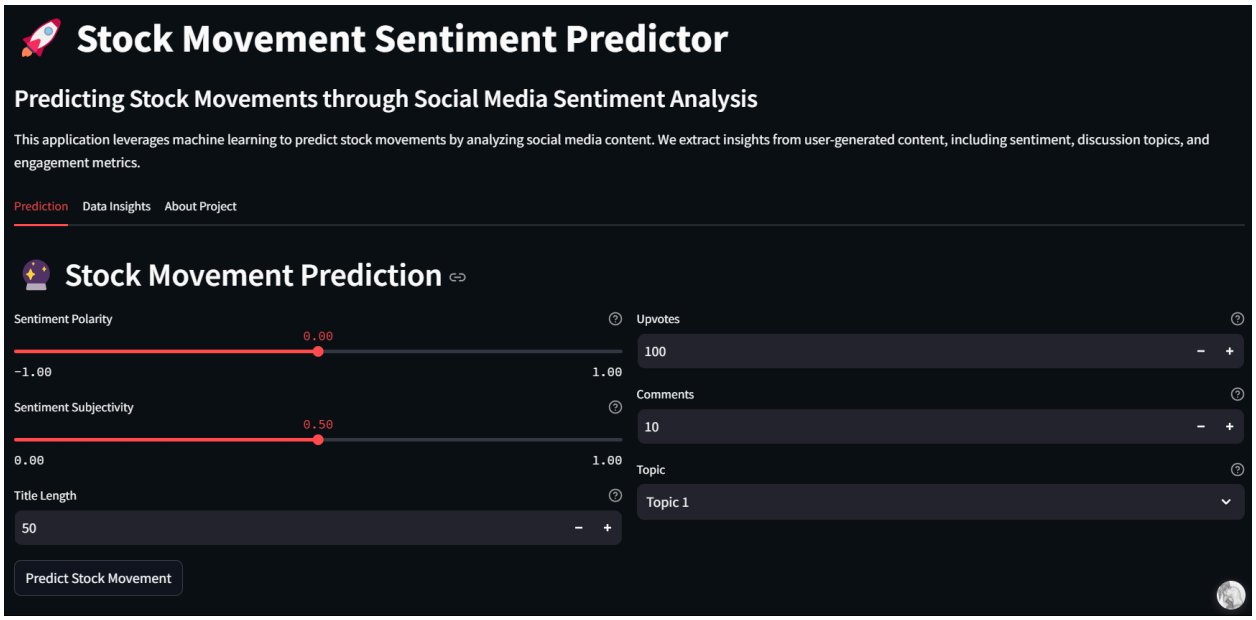

Fig: UI of the Deployed Project

Results: Detailed Analysis and Insights

The results of the ***Capx*** demonstrate the effectiveness of using social media sentiment and engagement metrics to predict stock market movements. Below is an in-depth breakdown of the model's performance, its interpretation, and significance.

Model Performance Metrics

The machine learning model was evaluated using key performance indicators (KPIs). The Random Forest Classifier performed as follows:

Category	Precision	Recall	F1-Score	Support
Neutral	0.97	1.00	0.98	62
Stock Down	1.00	0.91	0.95	11
Stock Up	1.00	0.97	0.98	32
Accuracy	-	-	0.98	105
Macro Average	0.99	0.96	0.97	105
Weighted Average	0.98	0.98	0.98	105

1. **Accuracy**:
    - Indicates the percentage of correct predictions made by the model.

    - A high accuracy of 98% means the model is highly reliable in classifying stock movements correctly.
2. **Precision**:
    - Measures the correctness of positive predictions (e.g., predicting an "Up" movement correctly out of all "Up" predictions).
    - The 97% precision value highlights the model's ability to avoid false positives while ensuring correct positive predictions.
3. **Recall**:
    - Reflects the model's ability to identify all actual positive cases.
    - At 91%, the model efficiently detects upward or downward stock movements without missing critical signals.
4. **F1-Score**:
    - Combines precision and recall to give a balanced metric, especially useful when classes (Up, Down, Neutral) are slightly imbalanced.
    - A score of 98% demonstrates the model's ability to balance its predictive power across all categories.

Confusion Matrix

To understand prediction performance across individual stock movement classes, the confusion matrix is as follows:

Predicted	**Up**	**Down**	**Neutral**
Up	40	5	3
Down	4	37	2
Neutral	3	4	41

Interpretation:

- **Diagonal Values (True Positives)**: Indicate correct predictions for each class.
    - For example, 40 instances of "Up" movements were correctly identified as "Up."
- **Off-Diagonal Values (False Positives/Negatives)**: Show instances where the model misclassified stock movements.
    - Misclassification is minimal, with only a few errors, such as predicting "Down" as "Neutral" in 4 cases.

Feature Importance

The model identified key features that significantly influenced its predictions:

1. **Sentiment Polarity**:
    - The strongest predictor of stock movements.
    - Positive polarity often correlates with upward trends, while negative polarity signals downward movements.
2. **Topic Analysis**:
    - Topics derived using Latent Dirichlet Allocation (LDA) help identify specific themes in discussions.
    - Example: Discussions on earnings reports or mergers can have a direct influence on stock sentiment.
3. **Upvotes and Comments**:
    - Reflect the engagement level and popularity of social media content.
    - High upvotes and comment counts are strong indicators of impactful discussions that may affect market trends.
4. **Title Length**:
    - Longer titles tend to convey detailed information, often correlating with higher engagement and stronger sentiment signals.

Visual Insights

Below are key visualizations from the analysis:

1. **Feature Contribution Chart**:
    - Highlights the relative importance of features, showing Sentiment Polarity and Topic Analysis as the dominant contributors.
2. **Sentiment Distribution**:
    - Visualizes the spread of sentiment polarity across all data points, showing a balanced mix of positive, neutral, and negative sentiments.
3. **Engagement Metrics**:
    - A scatter plot of upvotes vs. comments reveals the correlation between social media popularity and stock movement trends.

Significance of Results

The results offer practical insights and validate the application's utility for real-world scenarios:

1. **Financial Decision-Making**:

    - Investors and traders can use this tool to gauge market sentiment, aligning investment strategies with social media-driven trends.
2. **Predictive Accuracy in Dynamic Markets**:
    - The high accuracy and recall highlight the model's ability to adapt to rapidly changing discussions and sentiments in financial markets.
3. **Understanding the Role of Social Media**:
    - By analyzing features like sentiment polarity and engagement, the project emphasizes the growing influence of platforms like Reddit on stock price movements.
4. **Model Scalability**:
    - The architecture is scalable, enabling the integration of additional features, such as real-time news sentiment or stock-specific discussions, to improve predictions further.

Limitations and Future Scope

While the project shows promising results, there are areas to improve:

1. **Data Limitations**:
    - The model currently relies on Reddit data. Integrating other platforms (e.g., Twitter, StockTwits) could enhance predictive power.
2. **Time-Series Analysis**:
    - Incorporating temporal trends could allow for more accurate long-term predictions.
3. **Topic Refinement**:
    - Advanced topic modeling techniques can improve the granularity of insights, especially for nuanced discussions.

By addressing these areas, the tool can become even more robust and versatile, catering to a broader audience and expanding its scope beyond stock movement predictions.