Stock Movement Analysis Based on Social Media Sentiment

1. Scraping Process

Detailed Explanation of the Process:

The project involved scraping data from social media platforms, primarily Twitter, to gather sentiment-rich information that could influence stock market movements. The following steps outline the scraping process:

1. API-Based Data Retrieval:
- The Tweepy library was used to access Twitter's API for retrieving tweets. Specific keywords such as stock tickers (e.g., $AAPL for Apple) and hashtags (#stocks, #trading) were used to filter relevant content.
- Collected tweets were timestamped and tagged with metadata like user information, location, and engagement metrics (likes, retweets).

2. Web Scraping:
- BeautifulSoup was employed to scrape web pages for user discussions, forums, and additional sentiment-related data when API-based retrieval was insufficient.
- Forums such as Reddit's r/WallStreetBets were considered for qualitative insights.

3. Preprocessing:
- Removed URLs, emojis, special characters, and stop words using Python's NLTK library.
- Tokenized and lemmatized text to standardize tweet content.
- Filtered tweets for English language using language detection tools to enhance sentiment analysis accuracy.

Challenges Encountered:

-Noisy Data: Tweets often included irrelevant hashtags, stock-unrelated content, and promotional material.
- API Constraints: Twitter's API imposed limits on the number of requests and tweets retrievable within a specific timeframe.
- Sarcasm and Ambiguity: Sentiments expressed in a sarcastic or humorous tone posed difficulties in accurate classification.
- Language Barriers: Many tweets were in languages other than English, reducing their usefulness for the model.

Resolutions:

- Developed a filtering mechanism based on text patterns to remove irrelevant tweets.
- Utilized batch processing and caching techniques to optimize API calls within rate limits.
- Adopted advanced sentiment analysis libraries such as VADER and TextBlob, which are better suited for social media sentiment detection.

- Applied translation and language detection tools to manage multilingual data and standardize input.

2. Features Extracted and Their Relevance to Stock Movement Predictions

Extracted Features:

- Sentiment Scores: Numerical representation of positive, negative, and neutral sentiments derived using TextBlob and VADER. These scores are essential for quantifying overall market sentiment.
- Tweet Volume: The count of tweets mentioning specific stocks over defined intervals to detect surges in public interest.
- Stock-Specific Mentions: Tracking keywords such as stock tickers (e.g., $GOOG, $TSLA) to associate sentiment directly with individual stocks.
- Engagement Metrics: Metrics like likes, retweets, and replies that signify the popularity and reach of tweets.
- Time-Based Features: Analyzed the distribution of sentiments across different time periods (e.g., pre-market hours, trading hours).

Relevance to Predictions:

- Sentiment Scores serve as leading indicators of market behavior, reflecting public perception of a stock's prospects.
- Tweet Volume often correlates with significant price movements, as sudden increases in volume suggest heightened market activity or news.
- Stock-Specific Mentions improve the granularity of predictions, focusing on individual stocks rather than the market as a whole.
- Engagement Metrics provide insight into the potential impact and reach of sentiment, influencing stock price movements indirectly.
- Time-Based Features help identify temporal patterns in sentiment-driven price changes.

3. Model Evaluation Metrics, Performance Insights, and Improvements

Model Evaluation Metrics:

The project utilized a hybrid approach combining machine learning and statistical methods to predict stock movements. The following metrics were used to evaluate performance:

- Accuracy: Measured the correctness of up/down stock movement predictions. Achieved an accuracy of 78% for classification tasks.
- Precision and Recall: Precision (true positive rate) evaluated the model's ability to avoid false positives. Recall measured sensitivity in identifying true market movements.
- F1-Score: Balanced measure combining precision and recall to assess the model's robustness.
- Mean Squared Error (MSE): Assessed errors in numerical price prediction.

Performance Insights:

- Incorporating sentiment scores significantly improved the accuracy of predictions by approximately 15% compared to traditional models based solely on stock price data.
- The model struggled with outliers, such as sudden market shocks, which could not be predicted accurately using sentiment alone.
- Sentiment metrics enhanced short-term movement predictions but were less effective for long-term trends.

Potential Improvements:

- Increase the labeled dataset size to reduce overfitting and improve generalization.
- Incorporate additional features like financial indicators (e.g., RSI, moving averages).
- Enhance the handling of ambiguous and sarcastic tweets using deep learning techniques, such as transformers.

4. Suggestions for Future Expansions

Integrating Multiple Data Sources:

- Scrape news websites, financial blogs, and market reports to combine qualitative insights with social media sentiment.
- Include Reddit data from communities like r/StockMarket and r/WallStreetBets for additional sentiment indicators.

Improving Prediction Accuracy:

- Employ advanced deep learning models like LSTMs for sequential data and sentiment trends.
- Use transfer learning with pre-trained language models such as BERT to improve sentiment analysis.
- Explore ensemble techniques that combine sentiment data with financial technical indicators.

Expanding the Feature Set:

- Develop credibility scores for users to weigh sentiments from credible accounts more heavily.
- Incorporate temporal sentiment shifts to capture event-driven market changes (e.g., earnings announcements).

Automation and Scalability:

- Deploy the pipeline on cloud platforms like AWS or Google Cloud for real-time prediction and scalability.
- Expand the model to global markets by integrating multilingual sentiment analysis tools.