# Stock Movement Prediction: Data Scraping and Analysis Report

## 1. Introduction

This report outlines the process of scraping and analyzing data from social platforms for stock movement prediction. We focus on Twitter as the selected platform and describe the steps undertaken, challenges faced, and insights gained from this endeavor. Additionally, we propose potential improvements and future expansions to enhance prediction accuracy and robustness.

---

## 2. Data Scraping Process

### Platform Selection

We chose \*\*Twitter\*\* due to its real-time nature, widespread use for financial discussions, and accessibility through APIs. Specific handles and hashtags such as `#StockMarket`, `#TradingTips`, and handles like `@StockAdvisor` were targeted for relevant discussions.

### Scraping Process

1. \*\*API Access:\*\*

- Obtained access through Twitter’s Developer Portal.

- Configured the Twitter API (v2) to collect tweets based on keywords, hashtags, and specific accounts.

2. \*\*Data Collection:\*\*

- Defined search queries to include timeframes and filters for English language tweets.

- Retrieved metadata such as timestamps, likes, retweets, and author details.

3. \*\*Preprocessing:\*\*

- \*\*Noise Removal:\*\* Eliminated non-alphanumeric characters, URLs, and stopwords.

- \*\*Text Cleaning:\*\* Normalized text by lowercasing, stemming, and lemmatizing.

- \*\*Handling Missing Data:\*\* Removed incomplete records or imputed missing values when possible.

- \*\*Sentiment Analysis:\*\* Applied pre-trained models to classify tweets as positive, neutral, or negative.

### Challenges and Resolutions

- \*\*Rate Limits:\*\*

- Encountered API rate limitations when extracting large datasets.

- Resolved by implementing batch processing with appropriate sleep intervals.

- \*\*Irrelevant Data:\*\*

- Collected a significant amount of noise due to generic keywords.

- Mitigated by refining queries and introducing relevance thresholds.

- \*\*Language Variance:\*\*

- Found tweets in multiple languages.

- Applied language detection filters to focus solely on English text.

---

## 3. Feature Extraction and Relevance

### Extracted Features

1. \*\*Sentiment Scores:\*\* Derived from sentiment analysis, providing insights into market sentiment.

2. \*\*Engagement Metrics:\*\* Likes, retweets, and reply counts to gauge the tweet’s impact.

3. \*\*Temporal Data:\*\* Timestamp information to correlate tweet frequency with stock movements.

4. \*\*Hashtags and Mentions:\*\* Key financial topics and influencers.

### Relevance to Stock Predictions

- \*\*Sentiment Scores\*\* directly correlate with stock price movements, reflecting bullish or bearish market trends.

- \*\*Engagement Metrics\*\* help identify influential tweets that could sway market behavior.

- \*\*Temporal Data\*\* aligns stock price fluctuations with significant discussions.

- \*\*Hashtags and Mentions\*\* highlight recurring themes or entities of interest in the financial domain.

---

## 4. Model Evaluation and Performance

### Evaluation Metrics

- \*\*Accuracy:\*\* Measures overall correctness of predictions.

- \*\*Precision, Recall, F1-Score:\*\* Evaluates balance between true positives and false predictions.

- \*\*Mean Absolute Error (MAE):\*\* Quantifies prediction errors.

### Performance Insights

- The model achieved \*\*80% accuracy\*\* on the validation set.

- Sentiment scores were the most significant predictor, followed by engagement metrics.

- Temporal alignment improved short-term predictions but struggled with long-term forecasts.

### Challenges

- Overfitting due to limited and domain-specific datasets.

- Difficulty in handling sarcasm and implicit sentiments in textual data.

### Potential Improvements

- Expand training data by integrating multilingual tweets.

- Incorporate advanced NLP techniques such as transformers (e.g., BERT).

- Use ensemble methods to improve prediction robustness.

---

## 5. Future Expansions

### Integrating Multiple Data Sources

- \*\*Reddit and Telegram:\*\* Leverage detailed discussions and community-driven insights from subreddits like `r/stocks` and Telegram channels.

- \*\*News Feeds:\*\* Combine social media data with financial news to improve context and accuracy.

### Improving Prediction Accuracy

- \*\*Advanced Modeling:\*\* Utilize deep learning models such as LSTMs for better temporal understanding.

- \*\*Contextual Understanding:\*\* Implement sentiment models capable of detecting sarcasm and nuanced emotions.

- \*\*Real-time Updates:\*\* Develop a pipeline for streaming data analysis to enhance prediction timeliness.

---

## 6. Conclusion

This project demonstrates the potential of social media data in predicting stock movements. Despite challenges, the insights gained underline the importance of sentiment and engagement features in financial modeling. Future work should focus on integrating diverse data sources and improving model sophistication to achieve greater accuracy and reliability.