Pulse: Real-Time Market Sentiment Analyze

1. Introduction

The financial market reacts rapidly to public opinion shared across online platforms. Social discussions often influence stock prices before traditional news sources publish updates. The goal of this project, **Pulse: Real-Time Market Sentiment Analyzer**, is to analyze public stock-market-related discussions and extract actionable sentiment insights using Natural Language Processing (NLP).

This project goes beyond basic sentiment analysis by incorporating explainability, comparison views, and alert mechanisms to support decision-making.

---

2. Objectives

- Collect real-world stock market discussion data
- Preprocess and clean unstructured text
- Perform sentiment analysis using a transformer-based NLP model
- Visualize sentiment trends and comparisons
- Provide explainable insights using keyword analysis
- Build a professional dashboard for analysis

2.1 System Architecture Overview

The overall workflow of the Pulse system follows a clear end-to-end pipeline:

```
Reddit (Public JSON Data)
      ↓
Data Ingestion Scripts
      ↓
Text Cleaning & Preprocessing
      ↓
FinBERT Sentiment Analysis Model
```

```
        ↓
Sentiment Results Dataset
        ↓
Streamlit Dashboard (Alerts, Keywords, Comparison)
```

This modular architecture ensures scalability, maintainability, and clarity in data flow.

3. Data Collection

3.1 Data Source

- **Platform**: Reddit (public JSON endpoints)
- **Subreddits**:
  - r/stocks
  - r/wallstreetbets
  - r/investing
  - r/StockMarket

3.2 Data Type

- Post titles and self-text
- Publicly available data only (no authentication, no personal data)

3.3 Data Storage

- Raw data stored as CSV files
- Merged multi-source dataset created for processing

4. Data Preprocessing

The raw text data was cleaned using a custom preprocessing pipeline:

- Removal of URLs, mentions, hashtags, emojis, and special characters
- Conversion to lowercase
- Removal of extra spaces

This ensured the data was suitable for NLP model input.

5. Sentiment Analysis Model

5.1 Model Used

- **FinBERT (ProsusAI/finbert)**
- Transformer-based model trained specifically for financial text

5.2 Sentiment Classes

- Positive
- Neutral
- Negative

5.3 Handling Long Text

- Reddit posts exceeding model token limits were safely truncated
- This follows standard industry NLP practices

5.4 Output

- Sentiment label for each post
- Confidence score indicating prediction strength

---

6. Dashboard Development

A Streamlit-based interactive dashboard was developed to visualize and explore results.

6.1 Features

- Sentiment distribution visualization
- Sentiment confidence trend analysis
- Post-level exploration filtered by sentiment

---

7. Beyond Use-Case Enhancements

To exceed the basic project requirements, the following enhancements were implemented:

7.1 Sentiment Spike Alert System

- Monitors recent posts
- Detects abnormal increases in negative sentiment
- Displays alert or stability status

7.2 Keyword Explainability

- Extracts top keywords for each sentiment class
- Helps explain why sentiment is positive or negative
- Improves model transparency

7.3 Comparison View

- Positive vs Negative sentiment comparison
- Neutral sentiment count
- Keyword comparison across sentiments

These features transform the project into a decision-support system rather than a simple analysis tool.

8. Tools & Technologies

- **Programming Language**: Python
- **Data Processing**: Pandas, NumPy
- **NLP**: Hugging Face Transformers, FinBERT
- **Visualization**: Streamlit
- **Text Processing**: NLTK
- **Environment**: Visual Studio Code, Virtual Environment

9. Challenges Faced & Solutions

Challenge	Solution
Twitter API rate limits	Switched to Reddit public data
Long text exceeding model limits	Implemented safe truncation
Explainability requirement	Added keyword insights

9.1 Assumptions & Limitations

- The analysis is based on publicly available Reddit discussions only.
- Sentiment reflects public opinion and discussion tone, not actual stock price movements.
- Data volume depends on subreddit activity and post frequency.
- The system is intended for analytical and educational purposes, not financial advice.

9.2 Future Enhancements

- Integration with live financial news APIs.
- Advanced topic modeling using LDA or BERTopic.
- Time-based sentiment trend forecasting.
- Cloud deployment for scalability and continuous data ingestion.

10. Conclusion

This project successfully demonstrates an end-to-end NLP pipeline for market sentiment analysis using real-world data. By extending the system with alerts, explainability, and comparison views, the project exceeds the original use case and aligns closely with real-world analytical systems used in industry.