Aistox: A Sentiment-Aware Platform for Stock Market Prediction

Induj Tyagi
@Notnaut77
indujtyagi@gmail.com
GitHub: https://github.com/Notnaut77/Aistox-Sentiment-Aware-Stock-Market-Analysis-Platform

Abstract—**The stock market is a volatile environment where both quantitative indicators and qualitative sentiment impact asset prices. In this paper, we present Aistox, a sentiment-aware platform for stock market analysis and prediction. It incorporates Natural Language Processing (NLP), financial time-series data, and real-time social sentiment from Twitter, Reddit, and financial news outlets to predict stock movement directions. Our pipeline includes data ingestion, preprocessing, sentiment/emotion extraction, topic modeling, entity recognition, feature engineering, and model inference.**

I. Introduction

The financial markets are heavily influenced by investor sentiment, news, and macroeconomic factors. While traditional models rely primarily on structured financial indicators, the growing availability of unstructured data from news portals, social media, and forums demands integration of NLP-based techniques. Aistox aims to create a hybrid framework to predict stock movements by analyzing both financial and sentiment signals.

II. Related Work

Research such as Tetlock's work on news sentiment impact and Bollen et al.'s studies on Twitter sentiment analysis has shown strong correlation between sentiment and market behavior. Tools like FinBERT and BERTopic have proven useful in financial NLP tasks.

III. Factors Influencing Stock Prices

A. Fundamental Factors

These include Earnings Per Share (EPS), Revenue Growth, Balance Sheet Health, Dividend Yield, and Price-to-Earnings Ratio.

B. Technical Indicators

RSI, MACD, Bollinger Bands, Moving Averages, Support and Resistance levels are used for short-term trend detection.

C. Macroeconomic Indicators

Factors such as GDP Growth, Inflation, Interest Rates, Exchange Rates, FII flows, and Oil Prices contribute to market sentiment and liquidity.

D. Behavioral Factors

Social media trends, media headlines, and retail investor psychology (FOMO, herd behavior) play a role in short-term price actions.

IV. System Architecture Overview

The Aistox pipeline is broken into the following phases:

1) **Data Collection**: News, Tweets, Reddit Posts.
2) **Preprocessing**: Cleaning and tokenization.
3) **Sentiment Analysis**: Using VADER, TextBlob, FinBERT.
4) **Emotion Detection**: NRCLex, GoEmotions.
5) **NER and Topic Modeling**: spaCy and BERTopic.
6) **Feature Engineering and Labeling**: Mapping tickers and labeling price movements.
7) **Model Training**: Logistic Regression on TF-IDF and numerical features.
8) **Inference**: Real-time sentiment-based prediction.

V. Data Collection Modules

A. News Feed Parsing

We use `feedparser` and `newspaper3k` libraries to extract full articles from RSS feeds of economic outlets.

B. Twitter Integration

Using `tweepy` Streaming API with financial keywords to extract tweets. Tweets are stored in JSON format for analysis.

C. Reddit Extraction

Using `PRAW` and `Pushshift`, we pull data from subreddits like r/StockMarket and r/IndiaInvestments.

VI. Text Processing and NLP Modules

A. Preprocessing

Cleaning includes lowercasing, punctuation/emoji removal, stopword removal, and lemmatization.

B. Sentiment Classification

VADER and TextBlob compute sentiment polarity; FinBERT is used for domain-specific accuracy.

C. Emotion Detection

NRCLex is used to tag texts with emotions such as fear, greed, and anticipation.

D. NER and Topic Modeling

We apply spaCy for named entity extraction and BERTopic for clustering documents into meaningful topics.

VII. Feature Engineering and Labeling

A. Ticker Mapping

Entities identified by NER are matched with stock tickers using the Yahoo Finance API.

B. Labeling

We assign a binary label based on stock price movement: 1 if Close(t+1) ¿ Close(t), else 0.

C. Final Dataset

A merged dataset includes text, sentiment, emotion, topic ID, ticker, and target label for modeling.

VIII. Model Training and Evaluation

A. TF-IDF + Logistic Regression

TF-IDF vectors from cleaned text are combined with numerical features like sentiment scores and topic IDs. Logistic Regression is trained using scikit-learn.

B. Evaluation

We use accuracy, precision, recall, and F1-score to evaluate model performance.

IX. Real-Time Inference

A script called `predict_direction.py` accepts new text and performs preprocessing, feature extraction, and prediction, outputting `UP` or `DOWN` with confidence.

X. Limitations and Future Work

Despite promising results, real-time stock prediction remains uncertain due to market complexity. Future enhancements include:

- Neural models (e.g., BERT, XGBoost)
- Real-time dashboards with live data streaming
- Backtesting and portfolio simulation

XI. Conclusion

Aistox demonstrates how sentiment-aware systems can enhance financial forecasting. It fuses structured and unstructured data, NLP, and machine learning into a unified prediction pipeline.

References

References

[1] M. L. de Prado, *Advances in Financial Machine Learning*.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*.

[3] FinBERT: https://github.com/ProsusAI/finBERT

[4] VADER: https://github.com/cjhutto/vaderSentiment

[5] BERTopic: https://github.com/MaartenGr/BERTopic

[6] Aistox GitHub Repo: https://github.com/Notnaut77/Aistox-Sentiment-Aware-Stock-Market-Analysis-Platform