Sentiment Analysis on r/WallStreetBets and predicting the success of their Market Strategies

CSE342 - Statistical Machine Learning Course Project

Ritika Thakur
2022408
CSAI

Saksham Singh
2022434
CSE

Fig. 1. r/WallStreetBets

***Abstract*—This project aims to build a simple ML model that allows users to judge whether the market decisions in r/WallStreetBets - the biggest Reddit community for stock trading - are safe for an average person to invest in, and thus automates this process.**

I. Problem

In recent years, social media platforms such as Reddit's r/WallStreetBets have emerged as influential forums for discussions and opinions on stock investments. However, analyzing the vast amount of unstructured data generated by these communities, not to mention their colorful and profane jargon, and aggressive trading strategies, it presents a significant challenge for investors seeking to make informed decisions. The aim of this project is to develop an AI-driven solution that leverages sentiment analysis on r/WallStreetBets posts to identify potentially lucrative investment opportunities and subsequently predict the future value of these stocks.

II. Objectives

A. Sentiment Analysis

Develop a sentiment analysis model capable of accurately assessing the sentiment expressed in posts and comments on r/WallStreetBets. The model should be trained to differentiate between positive, negative, and neutral sentiments regarding specific stocks mentioned in the forum.

B. Stock Prediction Model

Design and implement a machine learning model for predicting the future value of stocks based on historical price data, market indicators, and sentiment analysis results. The model should be capable of generating accurate predictions over various time horizons.

C. Integration and Deployment

Integrate the sentiment analysis and stock prediction models into a unified system capable of processing real-time data from r/WallStreetBets and generating investment recommendations. Deploy the system in a scalable and accessible manner, ensuring seamless operation and user interaction.

III. Method

A. Sentiment Analysis Model

Utilize natural language processing (NLP) techniques such as word embeddings, recurrent neural networks (RNNs), or transformer-based models (e.g., BERT) to develop the sentiment analysis model. Train the model on a labeled dataset of r/WallStreetBets posts and comments, fine-tuning it to capture the nuances of financial sentiment.

For this to be automated, we require to implement web scraping techniques to collect data from r/WallStreetBets, focusing on posts containing discussions about specific stocks. This can be done using the PRAW library (Python Reddit API Wrapper) to extract relevant information such as stock symbols, sentiment scores, and associated timestamps from the top posts of the subreddit. Perform data preprocessing steps including text cleaning, tokenization, and normalization.

B. Stock Prediction Model

Employ machine learning algorithms such as linear regression, decision trees, or neural networks to build the stock prediction model. Incorporate features derived from historical price data, technical indicators, and sentiment analysis results as input variables. Train the model on historical stock data and evaluate its performance using appropriate metrics.

C. Integration and Deployment

Develop a web-based application or API that interfaces with the sentiment analysis and stock prediction models. Implement real-time data processing capabilities to ingest new posts from r/WallStreetBets and generate updated investment recommendations.

IV. Dataset

A. Reddit WallStreetBets Posts

We intend to train our Sentiment Analysis model on an open sourced dataset posted on Kaggle, featuring 53K posts from r/WallStreetBets, with attributes like their upvotes, comments, timestamp etc. All the individual entries in this dataset need to be labelled according to their ”emotions” manually for training.

B. Quandl

We intend to make use of Python stock API ’Quandl’ to fetch data regarding the stock market in a desired time span and create a dataset for the purpose of stock prediction. Quandl contains data in two formats - Time-series and Table. We intend to make heavy use of the table format to extract information regarding different stocks and training our model based on it.

V. Expected Results

- A robust sentiment analysis model capable of accurately gauging the sentiment expressed in r/WallStreetBets posts.
- An effective stock prediction model capable of generating reliable forecasts based on historical data and sentiment analysis insights.
- A user-friendly interface or API for accessing investment recommendations derived from the integrated system.
- Demonstrated improvements in investment decision-making outcomes compared to traditional approaches, validated through backtesting and performance evaluation.

VI. References

This project is inspired by Michael Reeves’ YouTube video titled ”I Gave My Goldfish $50,000 to Trade Stocks” which involves similar sentiment analysis techniques execute the task at hand.