

Data Analysis

**Internship Assignment Report**

**“Predicting market volatility and building short-term trading strategies using data from Reddit’s Wall Street Bets”**

### Author Mentor: Ranit Sen

### Ayush Sharma

### Pragati Devkar

### Manish Kumar

### Vivek Kumar

### Hrutwik Bothe Yasin Shah

Assignment Time: AUG-2021-SEP-2021

|  |  |
| --- | --- |
|  |  |
| **CONTENTS**  1. Introduction |  |
| 2. Cleaning |  |
| 3. Merging |  |
| 4. EDA |  |
| 4. Sentimental data analysis |  |
| 5. LSTM |  |
| 6. Deployment |  |

# *Introduction*

# The Game stop “short squeeze” dominated financial, and even primetime news headlines in January 2021. Stories of individual investors becoming millionaires overnight became virtually dime a dozen. Analyst pontification that the phenomenon signaled a shift in markets – a shift favoring individual investors for the first time in years.

# The primary driver of the game stop squeeze saga was the social news website reddit specifically the “subreddit’ r/ wall street bets. Wall-Streeter Bets (WBS) is an online hub for discussions surrounding speculative stock picks, and really grew into its own during g the short squeeze. A few users on WSB essentially filed the crazed buying of the GME or GameStop stock. Group like Wall-Streeters’Bets have shred in a new era of trading in which the general public has the power to cause a stock market disruption on its own

# Our project aims to carefully follow data scrapped from like Reddit’s WallstreetBEts as a means to predict market volatility and build a robust short-term trading strategy.

# In this paper, we described our experiments with the algorithms that aim to predict the future price of a security through analyzing posts on WSB. The long-term goal of the project would be to be build a trading bot that intelligently executes short-term trades based on insights gleaned from somewhat unconventional sources like WSB- aggregating results obtained by the algorithms described in this paper in order to make a decision.

# ***Cleaning***

# Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

# If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

# **Is this way we guys have done cleaning process.**

# 

# ***Merging***

# Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw data stored in multiple files, worksheets, or data tables, that you want to analyze all in one go.

# 

# In this our project **“Predicting market volatility and building short-term trading strategies using data from reddit’s wall street bets”** we have done CSV and JSON

# 

# ***EDA***

# **{Exploratory Data Analysis}**

# Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

# In this project we show the final target variable using the bar graph.

# 

# Relation among all the dependent and independent variables

# 

# 

# In this picture show the final tag positive and negative values.

# Data of variables are normally distributed using Distributed plot

# 

# 1) it seems that all the other variable like: close, High, low has similar line plot other than volume .

# 2) all the variables value seems to increase as per the time.

# 

# ***Sentimental data analysis***

# Sentiment analysis (or opinion mining) uses NLP to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

# 

# ***LSTM***

# 

# Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. ... LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

# ***Deployment***

# Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps.

# 

# ***Conclusion***

# Our most important conclusion is that some amount of correlation exists,and that WallStreetsBets posts do have an impact on the actual stock prices. The real challenge lies in bringing out this correlation, and we made an attempt at overcoming it by building a neural network.

# 

# ***THANK-YOU***