**Name: Shreyansh Kuntal**

**Enrollment No: 22324018**

**Branch: BS-MS Physics**

**Project: Stock Sentiment Analysis Using Machine Learning**

**PROBLEM STATEMENT:The project aims to develop a sentiment analysis model to predict the movement of stock prices based on textual data from news articles, social media posts, and other sources of financial news and opinions. By analyzing the sentiment expressed in these texts, the model will seek to uncover insights into investor sentiment and market sentiment, which can be valuable indicators for making informed trading decisions.**

**STEPS**

**Step1->Collecting Dataset**

**Step2->Performing sentiment Analysis to label the dataset**

**Step3->Finding buy, sell point, Sharp Ratio, maximum Drawdowns , no. of trades, Win ratio**

**Step4->Converting the textual dataset into vectors.**

**Step5->Training the model**

**Step6->Evaluation and Plotting**

**Step7->Hyper tuning and comparing results using different machine learning model**

**STEP 1:**

In this project the headlines are extracted from FINVIZ.

It has headlines from many sites and is more consistent. This consistency improves sentiment analysis.

The data for META and Tesla is taken from FINVIZ.

A graph with a line

Description automatically generated

A graph with blue lines

Description automatically generated

A graph of stock market

Description automatically generated

A graph of a person

Description automatically generated with medium confidence

Since FINVIZ has various restrictions involved on scrapping.

A screenshot of a computer

Description automatically generated

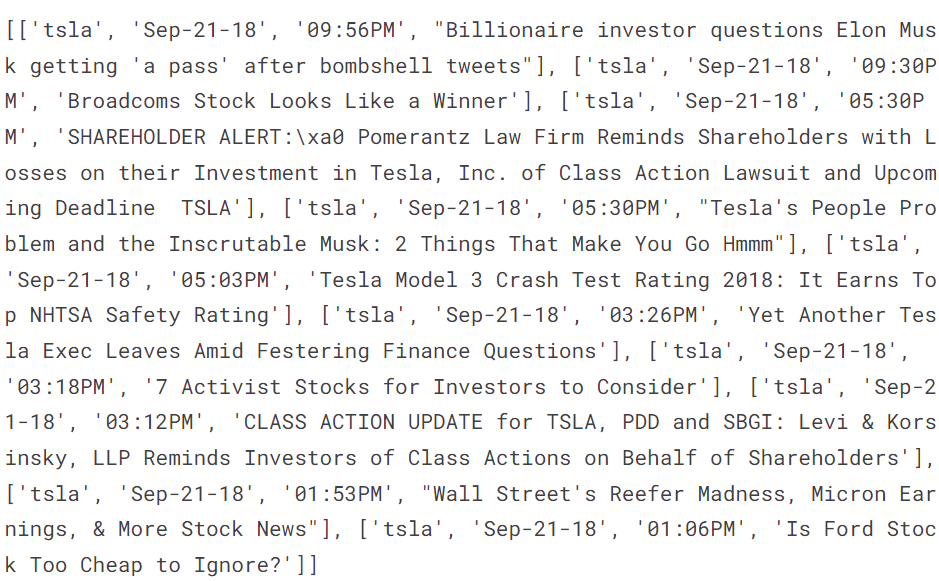
This was the type of data I was getting after scrapping. Therefore in this code HTML files for META and tesla at various points have been downloaded.

HTML files.

A white screen with black text

Description automatically generated

Scraping the textual Data.



**STEP 2:**

Performing sentiment analysis to label the data with neutral, negative and positive sentiments. As sentiment analysis is very context dependent we added some words with their score from our side.

new\_words = {

'crushes': 10,

'beats': 5,

'misses': -5,

'trouble': -10,

'falls': -100,

}

With these words included we updated the lexicon.

NLTK VADER is used for sentiment Analysis.

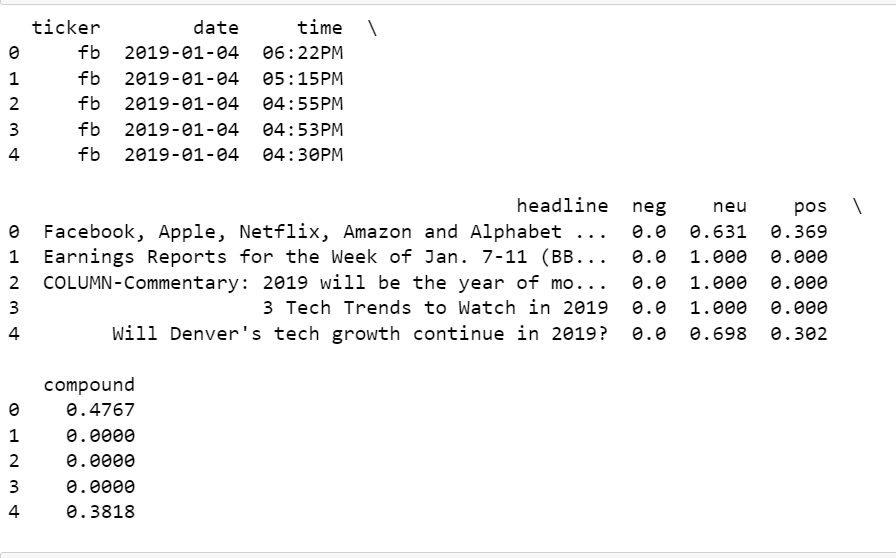
**What is Sentiment Analysis?**

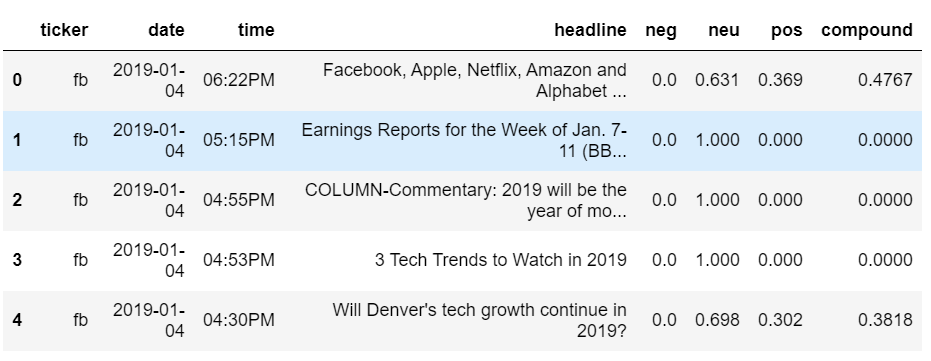
Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. It is a natural language processing (NLP) technique used to identify the emotional tone behind a body of text, such as customer feedback, social media posts, or reviews. The goal of sentiment analysis is to help organizations understand the opinions and emotions expressed in text data, which can be used to inform business decisions, improve customer service, and enhance marketing strategies.

**What is VADER?**

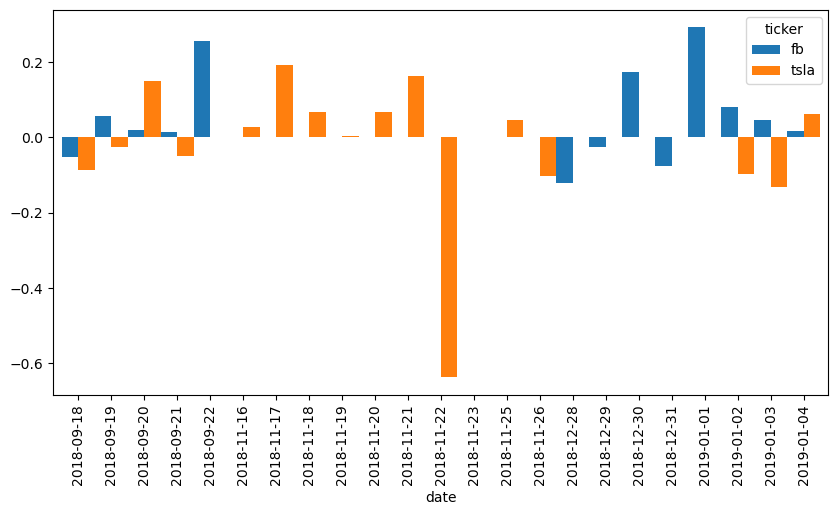
VADER stands for Valence Aware Dictionary and [Sentiment](https://www.analyticsvidhya.com/blog/2022/10/sentiment-analysis-using-vader/) Reasoner. It’s a tool used for sentiment analysis, which is basically a way to figure out if a piece of text is expressing positive, negative, or neutral emotions.

Using **Vader** we assigned polarity to each headline.

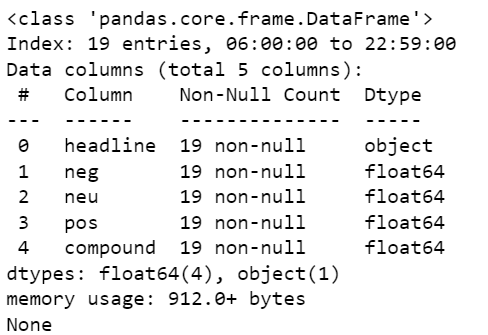




**Plotting the sentiments**



**Checking for only one single day**



**Plotting this single day data.**

A graph with numbers and a number of numbers

Description automatically generated with medium confidence

Labeling the dataset with neutral, negative and positive and then dividing the dataset of two stocks.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

### **STEP 3:**

### **1. Sharpe Ratio**

The Sharpe ratio measures the risk-adjusted return of an investment or trading strategy.

A mathematical equation with black text

Description automatically generated

**2. Maximum Drawdown**

Maximum drawdown measures the largest loss from a peak to a trough of a portfolio's value. It is typically expressed as a percentage:

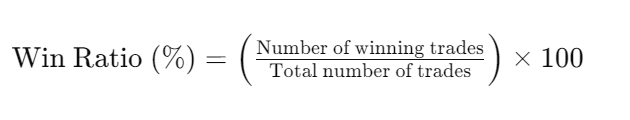
A math equation with black text

Description automatically generated

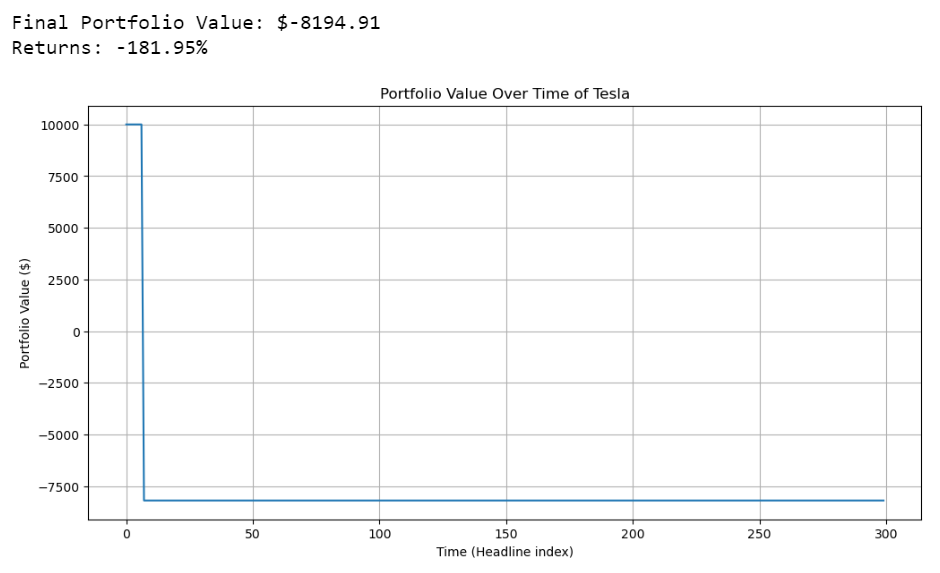
**3. Number of Trades Executed**

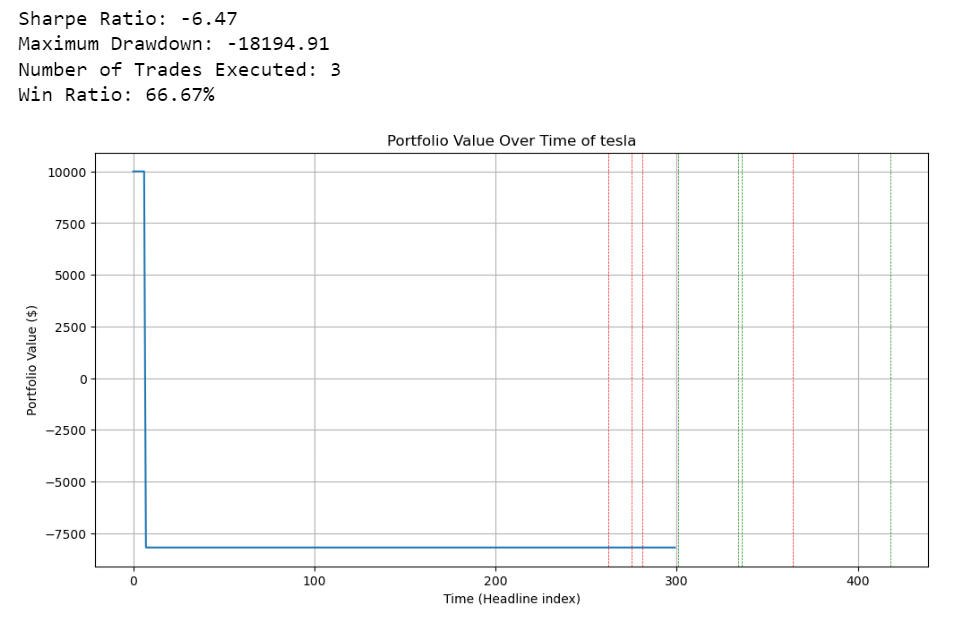
Simply counts the total number of buy and sell transactions executed during the trading period.

**4. Win Ratio**



Win ratio measures the percentage of winning trades out of total trades executed:





A graph with numbers and lines

Description automatically generated

A graph with numbers and lines

Description automatically generated

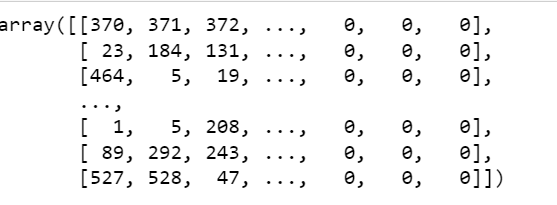
**STEP 4:**

For training the dataset the headlines were converted into vectors and the above labeling is used for supervised learning. To achieve this task the headline text was tokenized.

**TESLA:**

A black text on a white background

Description automatically generated



**META:**  
A close up of numbers

Description automatically generated

A number with numbers and dots

Description automatically generated with medium confidence

**STEP 5:**

**What is a neural network?**

A neural network is **a machine learning program, or model, that makes decisions in a manner similar to the human brain**, by using processes that mimic the way biological neurons work together to identify phenomena, weigh options and arrive at conclusions.

**Why neural network?**

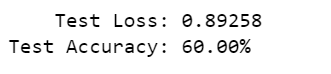
**Ability to Capture Non-linear Relationships**. Neural networks, especially deep neural networks (DNNs), can capture complex patterns and non-linear relationships in data. In NLP, language is inherently non-linear and nuanced, with dependencies that are often contextual and multi-layered. DNNs with multiple hidden layers can learn hierarchical representations of language, enabling them to model intricate linguistic features.

Training the neural network on both stocks’ datasets separately.

**STEP 6:**

To find better accuracy, the model was run for different number of epochs and number of dense layers and optimizer was also varied.

TESLA



A number with numbers and symbols

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated with medium confidence

**META**

A number with black text

Description automatically generated with medium confidence

A number with numbers on it

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated with medium confidence

**STEP 7:**

**TESLA**

Model: Neural Network

Accuracy: 0.6000

Precision: 0.7056

Recall: 0.6000

F1-score: 0.5242

Model: Logistic Regression

Accuracy: 43.33%

Precision: 42.44%

Recall: 43.33%

F1 Score: 41.80%

Model: Support Vector Classifier

Accuracy: 46.67%

Precision: 75.11%

Recall: 46.67%

F1 Score: 29.70%

Model: Random Forest

Accuracy: 48.33%

Precision: 49.82%

Recall: 48.33%

F1 Score: 43.81%

Model: Gradient Boosting

Accuracy: 48.33%

Precision: 47.28%

Recall: 48.33%

F1 Score: 45.23%

Model: K-Nearest Neighbors

Accuracy: 51.67%

Precision: 52.04%

Recall: 51.67%

F1 Score: 48.93%

Model: Decision Tree

Accuracy: 40.00%

Precision: 40.53%

Recall: 40.00%

F1 Score: 40.10%

BEST MODEL AND PARAMETERS

SVC (C=1, gamma=1)

{'C': 1, 'gamma': 1, 'kernel': 'rbf'}

**META**

Model: Neural Network

Accuracy: 0.6000

Precision: 0.7056

Recall: 0.6000

F1-score: 0.5242

Model: Logistic Regression

Accuracy: 43.33%

Precision: 42.44%

Recall: 43.33%

F1 Score: 41.80%

Model: Support Vector Classifier

Accuracy: 46.67%

Precision: 75.11%

Recall: 46.67%

F1 Score: 29.70%

Model: Random Forest

Accuracy: 46.67%

Precision: 46.67%

Recall: 46.67%

F1 Score: 39.85%

Model: Gradient Boosting

Accuracy: 48.33%

Precision: 47.28%

Recall: 48.33%

F1 Score: 45.23%

Model: K-Nearest Neighbors

Accuracy: 51.67%

Precision: 52.04%

Recall: 51.67%

F1 Score: 48.93%

Model: Decision Tree

Accuracy: 40.00%

Precision: 39.75%

Recall: 40.00%

F1 Score: 39.82%