Title: Stock Market Sentiment Analysis

First name	Last Name	IIT Email
Akshatha	Ganji	aganji1@hawk.iit.edu
Joel	Miranda	jmiranda11@hawk.iit.edu
Sree Ram Charan	Pammi	spammi@hawk.iit.edu
Varshitha	Chikkegowdanadoddi Somashekar	vchikkegowdanadoddi@hawk.iit.edu

1. Introduction

On social media, information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Twitter, a social media platform, has received a lot of attention from researchers in recent times. Twitter is a micro-blogging application that allows users to follow and comment on other user's thoughts or share their opinions in real time.

More than a million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers. Each tweet is 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions. Sentiment analysis of twitter data and sentiment classification is the task of judging opinion in a piece of text as positive, negative, or neutral. In this project a method for predicting stock prices is developed using Twitter tweets about various companies. Sentiment analysis of the collected tweets is used as a prediction model for finding and analyzing correlation between contents of news articles and stock prices and then making predictions for future prices will be developed by using machine learning.

Sentiment analysis of "Tweets about the Top NASDAQ Companies from 2015 to 2020" dataset from Kaggle and examining their effects on stock market changes. Even headlines from different news sites can also be included. Awareness and click generation are important roles for business news headlines as well. The effect of news and tweets on the sentiment of the stock market can be understood. Also, comparison of a few Natural Language Processing (NLP) algorithms will be considered.

2. Data Sets

Tweets about the Top NASDAQ Companies from 2015 to 2020
https://www.kaggle.com/code/tommyupton/twitter-stock-market-sentiment-analysis/input

This dataset as a part of the paper published in the 2020 IEEE International Conference on Big Data under the 6th Special Session on Intelligent Data Mining track, is created to determine possible speculators and influencers in a stock market. Although we used both tweet data and companies' market data in our project, we thought that it is a better choice to split our datasets into two parts while sharing in Kaggle. This dataset is helpful for those interested in tweets that are written about Amazon, Apple, Google, Microsoft, and Tesla by using their appropriate share tickers.

This dataset contains over 3 million unique tweets with their information such as tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets of tweets matched with the related company.

Tweets are collected from Twitter by a parsing script that is based on Selenium.

3. Research Problems

Stock exchange is a subject that is highly affected by economic, social, and political factors. There are several factors e.g. external factors or internal factors which can affect and move the stock market. Stock prices rise and fall every second due to variations in supply and demand. Various Data mining techniques are frequently involved to solve this problem. But it has been observed it takes longer time, not accurate and complex routines involved in understanding the sentiment of stock market. Predicting the sentiment of market can give lots of input in predicting stock prices and do investment accordingly.

4. Potential Solutions

Techniques using machine learning will give more accurate, precise and simple way to solve such issues related to stock and market sentiment and prices. Positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. A prediction model for finding and analyzing correlation between contents of tweets and stock prices and then making predictions for future prices can be developed by using machine learning

This project concentrates on using Python packages such as NLTK (Natural Language Tool Kit) , Pandas, NumPy, Scikit-Learn, Matplotlib and TensorFlow.

5. Evaluation Plans

The following are the evaluation plan

- Collect Dataset from Kaggle
- Preprocess using NLTK
- Use Count Vectorizer for word embedding
- Implement Classifiers are - Random Forest, Multinomial Naive Bayes, Logistic Regression, K-Nearest Neighbors and Support vector Classifier

- Evaluating all the models using the below techniques
  - Get the test dataset and tokenize the news titles by words
  - Cleanup by removing the stop-words
  - Represent test data using word-cloud
  - Stemming and joining the words
  - Build a corpus of test data and visualize using word-cloud
  - Test the model using the classifiers mentioned in the previous step.
  - Finally, confusion matrices will be generated to visualize and summarize the performance of all the classification algorithms.

The metrics used for the evaluation are:

- Bullish sentiment (Measures the performance based on positive sentiments)
- Bearish sentiment (Measures the performance based on negative sentiments)
- Neutral sentiment (Performance measurement by neutral sentiments)
- Current market trend (Captures current trends of the market based on sentiments)

The performance metrics measured for each classification algorithms are:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC Curve

These metrics together provide a comprehensive understanding of the model's ability to correctly classify sentiments as bullish, bearish, or neutral, and its effectiveness in capturing the current market trend based on sentiment analysis.

6. Expected Outcomes

We can evaluate how the sentiment is assigned to the trades, a better understanding of the tweet sentiment might lead to more interesting results. Looking at stock trends versus the market as a benchmark would also add value to the analysis. It will also help us understand how to produce better machine-learning models to solve problems with the help of solving the problem statements mentioned above.