Assignment 2

CPSC 488: NATURAL LANGUAGE PROCESSING
10/14/2024
DR. CHRISTOPHER RYU
CALIFORNIA STATE UNIVERSITY, FULLERTON (CSUF)

Assignment Information

This assignment can be completed individually or by a team with up to 5 members.
Total score: 100
Due date: 12/01/2024 11:59PM
Members: Hammad Sheikh, Matthew Do, Ryan Avancena
Work distribution: Equal

Abstract - Machine learning (ML) and Artificial Intelligence (AI) worlds are experiences tremendous growth. Their applications are uncountable. One realm of applications for ML and AI is Natural Language Processing (NLP). In this aspect, there are various classification models that assist with text classification and processing. Our goal for this assignment is to utilize NLP techniques in analyzing news data for ETF FNGU to build a ML model that would maximize our investment returns. The stock data can be obtained via the Python libraries that Dr. Ryu [1] defined. News articles will be scraped from the web.
Keywords - Stock Trading, ETF, FNGU, Natural Language Processing, Machine Learning

Contents

1 Introduction

There is a lot of stock data available in the world. All the stocks are impacted by the environment and the economical status of the world. This information is shared via news articles, which are in text format. We need to able to analyze news data, identify patterns and sentiment, and utilize it to predict impact on stocks. However, due to the sheer size of data available and the amount of convolution, it is not an easy task for a human. If we can utilize the computing power and technology available in computers, we may be able to make an impact. This may not have been possible a few decades ago. However, with the expansion of ML and AI technologies, numerous applications become available. With the use of NLP, we can perform text classification and sentiment analysis. Without generalizing too much, let us look at an ETF, ticker FNGU, with the goal of applying existing text classification and sentiment analysis algorithms on news and stock data, and analyzing their results for performance and accuracy to maximize our investment returns.

2 Approach

The generalized approach is shown below in Figure 1. (Data Acquisition → Data Pre-Processing → Data Classification and Sentiment Analysis → Model Development for Investment ↔ Parameter and Feature Adjustment to Maximize Investment Returns.) Further details will be discussed in the following sections.

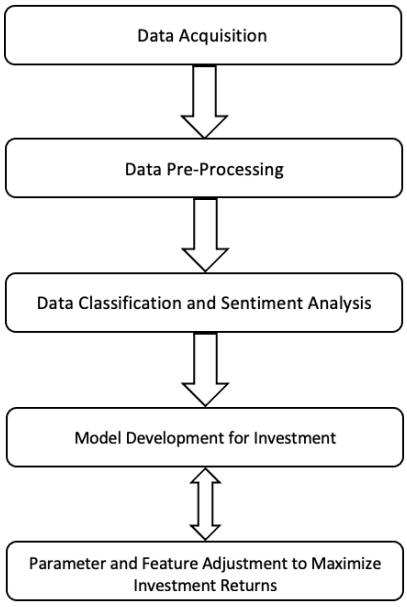

Figure 1: Generalized Approach

We will be relying on existing libraries as much as possible. This will help to focus on the classification and sentiment analysis task instead of building code from scratch for things that already exist. For example, train_test_split from sklearn.model_selection [2] will be utilized to split the data into training and testing datasets.

2.1 Data Acquisition

The first step in data processing, regardless of whether it is for our assignment or any other project, is data acquisition. We need to have a dataset that can be used to train need based models.

Our assignment needs two main datasets; one for stock information and one for news articles. For the ETF information, we utilized Python library yfinance [3] to download the data and stored it locally in JSON files. For the news articles, we utilized get_news() function of the yfinance library [3] to grab the URLs and then grabbed the articles text and information via HTML parser, modified from Dr. Ryu's [1] code. All the information grabbed is stored in appropriate JSON files to be utilized in model development. This enables us to have a defined set of data, which we can preprocess and utilize in our model development of sentiment analysis and investment returns.

Figure 2 shows a sample snapshot of FNGU market data acquired from yfinance [3], and Figure 3 shows a sample snapshot of news data scraped from the web via our HTML scraper.

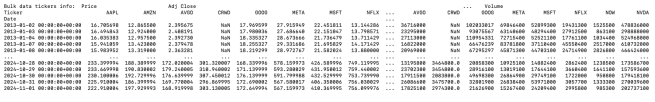

Figure 2: Sample snapshot of FNGU market data

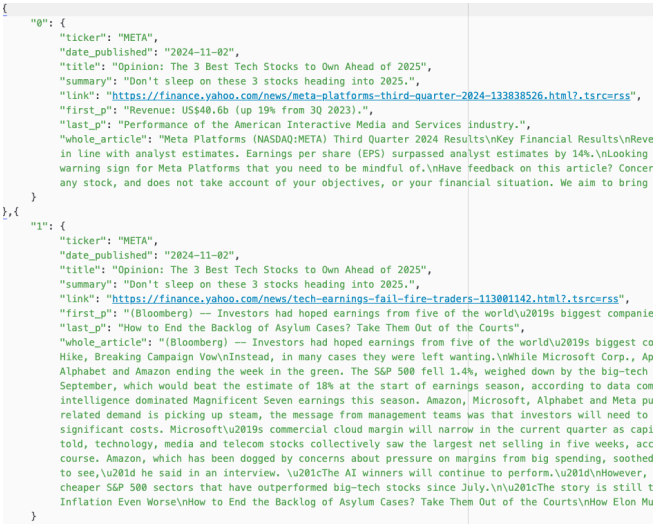

Figure 3: Sample snapshot of scraped News data

2.2 Data Pre-Processing

2.3 Data Classification and Sentiment Analysis

Model/Evaluation Metric	Accuracy	Sensitivity or Recall	Specificity	Precision	Harmonic mean
Bernoulli Naïve Bayes	86.18%	21.89%	94.88%	36.65%	27.41%
Logistic Regression	88.30%	17.63%	97.87%	52.78%	26.43%
Default MLP	88.74%	41.56%	94.55%	50.79%	45.71%
Adjusted MLP	90.05%	34.14%	97.61%	65.95%	44.99%

Table 1: Classifiers' Evaluation Metrics

2.4 Model Development for Investment

2.5 Parameter and Feature Adjustment to Maximize Investment Returns

3 Conclusion and Future Work

4 Acknowledgements

I would like to acknowledge Dr. Jin [4] for ML, classification and probability distribution concepts, Professor Avery [5] for MLP concepts,

References

[1] C. Ryu, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.

[2] scikit learn, "train_test_split," https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

[3] Y. Finance, "yfinance," https://pypi.org/project/yfinance/.

[4] R. Jin, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.

[5] K. Avery, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.