IE 6318 – Project Assignment

Name – Ankit Ranjan

Student ID – 1001832327

Guided by – Dr. Shouyi Wang

Title: Exploring alternative data and use it in conjunction with traditional data to predict stock prices with trading strategies

Abstract:

The stock exchange has been around for centuries, and consequently stocks price prediction is an age-old challenging to tackle. There exist many strategies that use traditional data like historical data to help model their predictions. A key factor in these models, that is omitted in predicting the stock prices is alternative data which enhances the confidence of predictions. There are many types of alternative data like data from social media web scraping, web traffic, credit card and point-of-sale, geolocation and satellite imagery, that may be used. However, the ones that are being explored in this project are the Financial News delivery services and a sentiment analysis will be performed on them. The services include Google News headlines, headlines from Financial data vendor like Intrinio, and StockTwits posts.

Introduction:

The aim of stock market prediction is to predict the future movement of a financial exchange's stock value. Investors would be able to turn a profit if they can accurately predict the price change in the share. Predicting how the stock market will move is one of the most difficult tasks because there are so many variables to consider, such as interest rates, politics, and economic development. Since stock investment is a major financial market activity, a lack of accurate knowledge and comprehensive information will inevitably result in a loss of investment.

Stock market prediction methods are divided into two main categories: technical and fundamental analysis.

Fundamental and technical analysis are two major schools of thought when it comes to approaching the markets yet are at opposite ends of the spectrum. Investors and traders use both to research and forecast future stock prices. Like any investment strategy or philosophy, both have advocates and adversaries.

Fundamental Analysis evaluates stocks by attempting to measure their intrinsic value. Fundamental analysts study everything from the overall economy and industry conditions to the financial strength and management of individual companies. Earnings, expenses, assets and liabilities all come under scrutiny for fundamental analysts.

Technical Analysis differs from fundamental analysis, in that traders attempt to identify opportunities by looking at statistical trends, such as movements in a stock's price and volume. The core assumption is that all known fundamentals are factored into price, thus there is no need to pay close attention to them. Technical analysts do not attempt to measure a security's intrinsic value. Instead, they use stock charts to identify patterns and trends that suggest what a stock will do in the future.

Alternative data is being used by fundamental and quantitative institutional investors to create innovative sources of alpha. Examples of alternative data include: Geolocation (foot traffic), Credit card transactions, Email receipts, Point-of-sale transactions, Web site usage, Mobile App or App Store analytics, Obscure city hall records, Satellite images, Social media posts, Online browsing activity, Shipping container receipts, Product reviews, Price trackers, Shipping trackers, Internet activity and quality data. However, the ones that are being explored in this project are the Financial News delivery services and a sentiment analysis will be performed on them. The services include Google News headlines, headlines from Financial data vendor like Intrinio, and StockTwits (social media platform designed for sharing ideas between investors, traders, and entrepreneurs) posts.

Sentiment analysis is a common method which is increasingly used to assess the feelings of social media users towards a subject. According to the efficient market hypothesis, all past information is reflected in stock prices and new information is instantaneously absorbed in determining future stock prices. Hence, prompt extraction of positive or negative sentiments from news (alternative data) is very important for investment decision-making by traders, portfolio managers and investors. Sentiment analysis models can provide an efficient method for extracting actionable signals from the news. However, financial sentiment analysis is challenging due to domain-specific language and unavailability of large, labeled datasets. General sentiment analysis models are ineffective when applied to specific domains such as finance.

Related work:

In 2018, a study by Sahar Sohangir et al., titled 'Big Data: Deep Learning for financial sentiment analysis' (where they got permission to obtain data from StockTwits Inc. directly), they found that, among different common Deep Learning methods in sentiment analysis, only convolutional neural network outperforms logistic regression. The accuracy of convolutional neural networks, in comparison to the other models, was considerably better. There are some people in the financial social network who can correctly predict the stock market. By using CNN to predict their sentiment they attempted to predict future market movement.

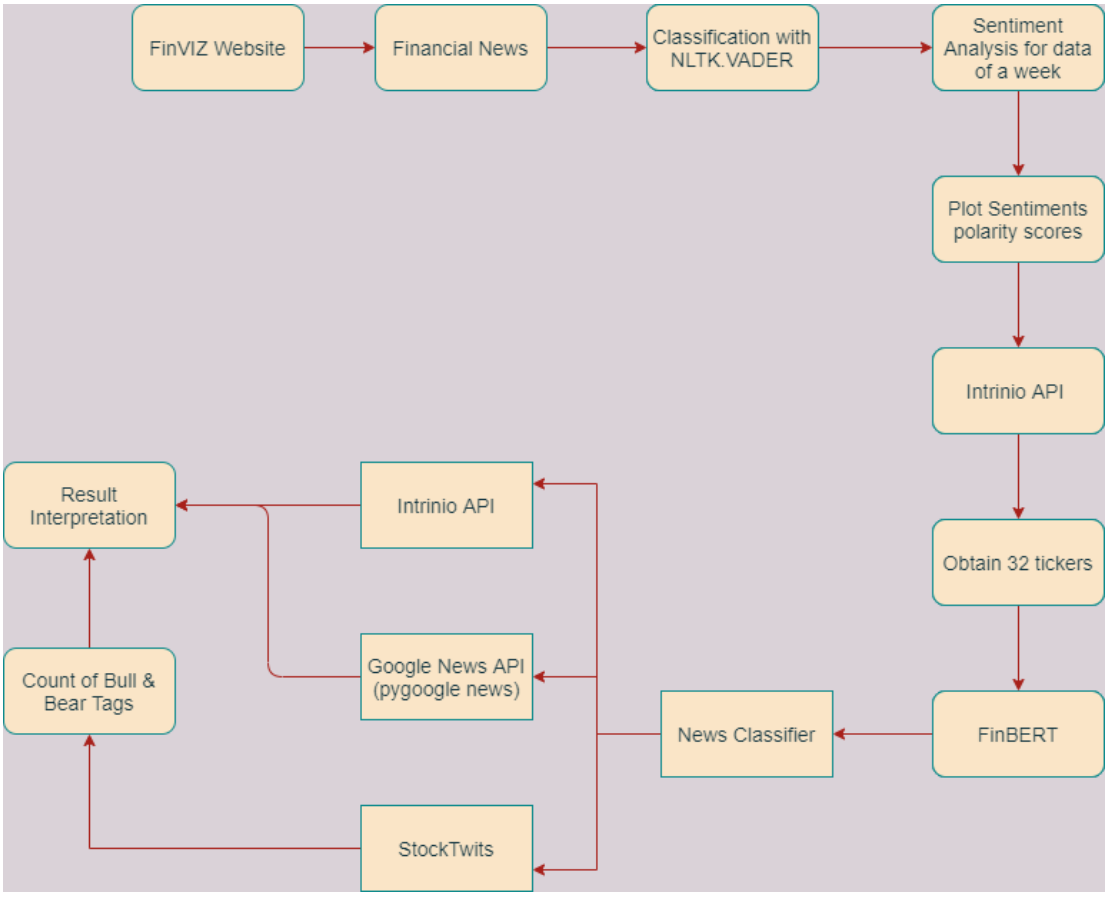
Another summary was obtained in 2020 by Aaryan Gupta et al. where they found that in a study by Joshi et al. (2016) compared three ML algorithms and observed that random forest (RF) and SVMs performed better than NB. Renault (2019) used StockTwits (a platform where people share ideas about the stock market) as a data source and applied five algorithms, namely NB, a maximum entropy method, a linear SVM, an RF, and a multilayer perceptron and concluded that the maximum entropy and linear SVM methods gave the best results. Over the years, researchers have combined deep learning methods with traditional machine learning techniques (e.g., construction of sentiment lexicon), thus obtaining more promising results (Yang et al. 2020).

Next, in a study by Anita Yadav et al., their experiments have demonstrated that it is possible to improve upon the traditional techniques for determining sentiments using sentiments indicators. Both hybrid approach and noun-verb approach showed better results than Turney's approach, with noun-verb being found best for the financial texts.

In 2018, the paper on the current SOTA NLP model for tasks like Sentiment Analysis (classification), Named Entity Recognition, and Q&A (the software receives a question and is required to mark the answer. Using BERT, the model marks the beginning and the end of the answer.) was published. It was named BERT (Bidirectional, Encoder, Representations from Transformers) and it is Google's neural network-based lates model for NLP pre-training. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Consequently, FinBERT was created by pre-training BERT on three datasets: TRC2-financial (a subset of Reuters' TRC2), Financial PhraseBank and FiQA Sentiment, to make it applicable for financial context. The FinBERT model beat the previous SOTA (LSTM model (by 15 %) for classification task even with a small training set of 500 examples. This was all the more impressive given that the dataset Financial PhraseBank suffered from label imbalance.

Flow Chart:



Definitions:

Ticker – A ticker symbol or stock symbol is an abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market. A stock symbol may consist of letters, numbers or a combination of both.

Transformer – A Transformer is a deep learning model introduced in 2017 that utilizes the mechanism of attention, weighing the influence of different parts of the input data. It is used primarily in the field of natural language processing (NLP).

Python Toolboxes:

numpy, pandas, matplotlib for array processing and visualization.

BeautifulSoup to process HTML files and extract important features from it (news headlines).

nltk to carry out sentiment analysis using the VADER module.

transformer to import the pipeline module, which helps import FinBERT model to carry out sentiment classification.

intrinio_sdk API to gather tickers and their corresponding headlines.

pygooglenews API to find headlines of tickers obtained from intrinio_sdk.

sklearn to compute the accuracy of classification of pre-tagged StockTwits messages.

Exploratory Data Analysis:

Parameter Name	Unit	Description
Input Parameter:		
Financial news / post	String	It is the headline / post to be analyzed by the sentiment analyzer (BERT / NLTK Vader)
Output Parameters:		
Sentiment	String	The output will be the sentiment (i.e., negative, neutral or positive) with an associated score.

For the NLTK Vader testing part of this project, 5 HTML files containing financial headlines were downloaded from FINVIZ website and explored offline so as to not send unnecessary requests to the FINVIZ website.

For the BERT testing part of this project, by providing an API key obtained by creating an account on their website, 32 tickers were extracted from the Intrinio API (intrinio_sdk). Using these tickers, their corresponding news headlines were extracted using the Intrinio API (intrinio_sdk) and Google News API (pygooglenews), and StockTwits' users' messages (AKA posts).

For both the above tests, an array of string input was passed from a data frame to compute the sentiment of the strings as the output.

Preprocessing:

For the NLTK Vader testing part of this project, each HTML file was converted into a BeautifulSoup object (a web scraping tool) and each HTML file contained a 'news-table' table. Inside each news-table, each row tag (<tr>) was entered and within that were the <a> and <td> tags which contained the URL and headline, respectively. The headlines were thus extracted and stored in a data frame. Additionally, the time of publishing of the headlines were also stored in the data frame. The NLTK Vader lexicon was updated with a few financial words to reflect the financial jargon more appropriately. Finally, these headlines were used as input to NLTK's Vader module, and the negative, neutral, positive and compound scores were computed for each headline.

For the FinBERT testing part of this project, after the 32 tickers were extracted from the Intrinio API, the news headlines' titles, the date and time of publishing and corresponding tickers were extracted and stored in a data frame. Next, from the Google News API, the tickers extracted from Intrinio API were used to extract headlines of the last 3 days and along with the publishing date and URL, they were added to a data frame. Next, for StockTwits' users messages, the json pages for each ticker which had messages on them were obtained from pages in the form of the URL: 'https://api.stocktwits.com/api/2/streams/symbol/{ticker}.json' where any 'ticker' could be substituted in the URL format to obtain the recent messages on that stock symbol. In the JSON URLs: recent messages, their labelled sentiment i.e., bull tag / bear tag / None tag (no assigned sentiment) were embedded deep in dictionaries. These attributes were extracted and stored in a data frame.

The ProsusAI/FinBERT version of the FinBERT model was used in classification. It was obtained from HuggingFace's Transformers (which is bult on top of PyTorch).

Thus, the headlines and messages are in a desired form (strings) for the input data for the machine learning models (NLTK Vader, transformers) is obtained.

Number of headlines by tickers obtained from Intrinio API:

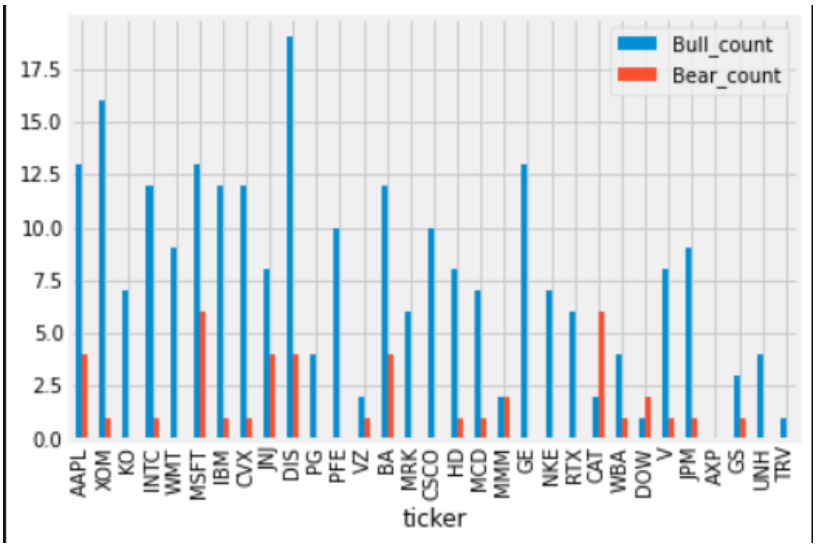
	tickers
MCD	100
MRK	100
TRV	100
UNH	100
DOW	100
VZ	100
XOM	100
PFE	100
RTX	100
INTC	100
WBA	100
JPM	100
WMT	100
AXP	100
HD	100
AAPL	100
MSFT	100
IBM	100
NKE	100
JNJ	100
GS	100
UTX	100
CAT	100
DIS	100
GE	100
BA	100
MMM	100
CSCO	100
V	100
KO	100
PG	100
CVX	100

Number of headlines by tickers obtained from Google News API:

	ticker
MSFT	100
JNJ	100
UNH	100
DOW	100
XOM	100
PFE	100
INTC	100
WBA	100
JPM	100
WMT	100
HD	100
AAPL	100
IBM	100
CVX	100
MMM	100
CSCO	100
PG	100
UTX	100
CAT	100
GE	100
V	100
GS	100
DIS	97
AXP	92
NKE	88
MRK	88
MCD	85
BA	81
RTX	70
KO	56
VZ	46

Tagged recent messages by tickers obtained from StockTwits JSON URL:

	ticker	Bull_count	Bear_count	Total_messages
0	AAPL	13	4	17
1	XOM	16	1	17
2	KO	7	0	7
3	INTC	12	1	13
4	WMT	9	0	9
5	MSFT	13	6	19
6	IBM	12	1	13
7	CVX	12	1	13
8	JNJ	8	4	12
9	DIS	19	4	23
10	PG	4	0	4
11	PFE	10	0	10
12	VZ	2	1	3
13	BA	12	4	16
14	MRK	6	0	6
15	CSCO	10	0	10
16	HD	8	1	9
17	MCD	7	1	8
18	MMM	2	2	4
19	GE	13	0	13
20	NKE	7	0	7
21	RTX	6	0	6
22	CAT	2	6	8
23	WBA	4	1	5
24	DOW	1	2	3
25	V	8	1	9
26	JPM	9	1	10
27	AXP	0	0	0
28	GS	3	1	4
29	UNH	4	0	4
30	TRV	1	0	1

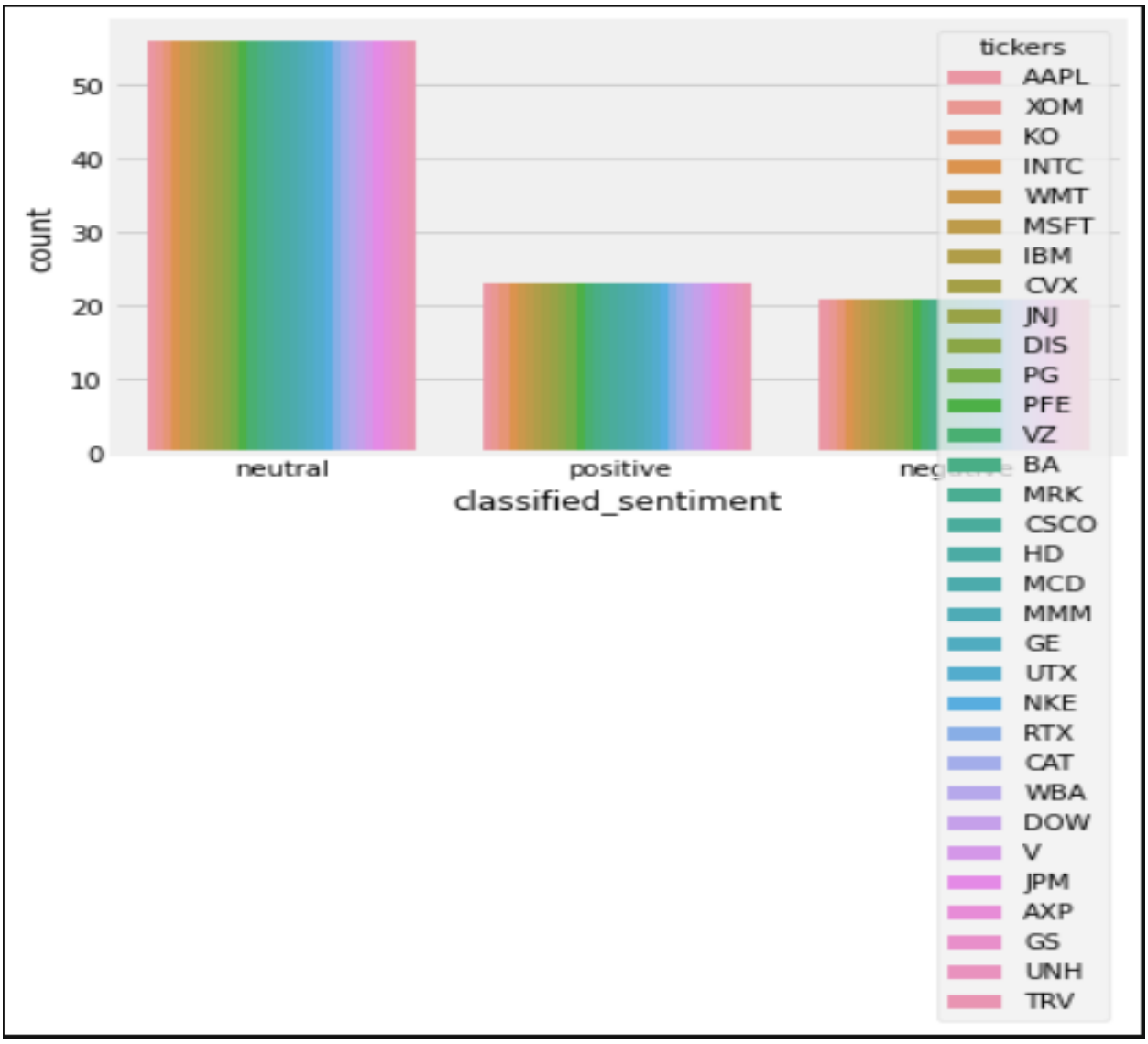


Data visualization:

For headlines obtained from FINVIZ and analyzed by NLTK Vader:

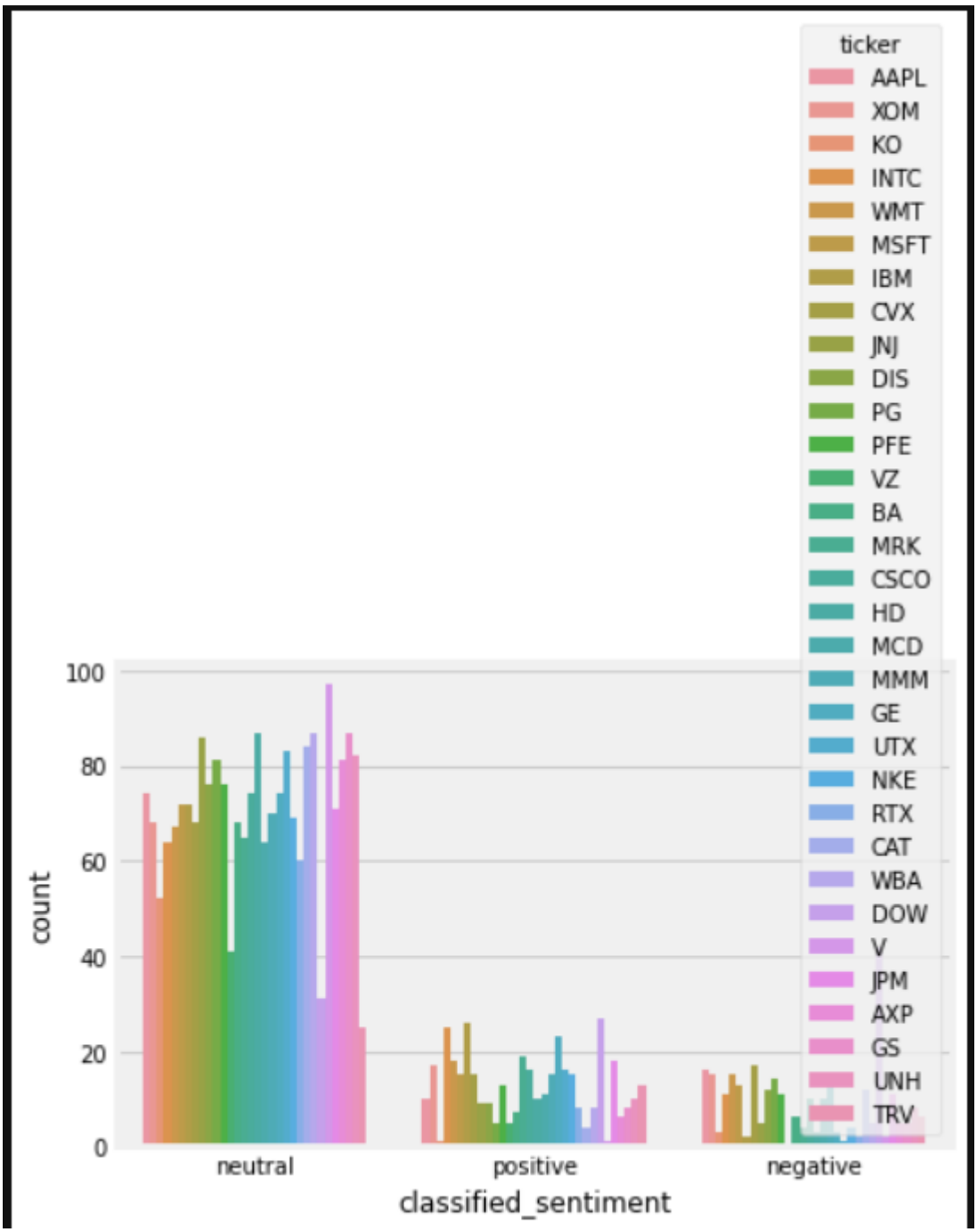
The first 40 headlines were plotted time wise (in date order) with their sentiment breakdown.

For headlines obtained from Intrinio API and analyzed by FinBERT:



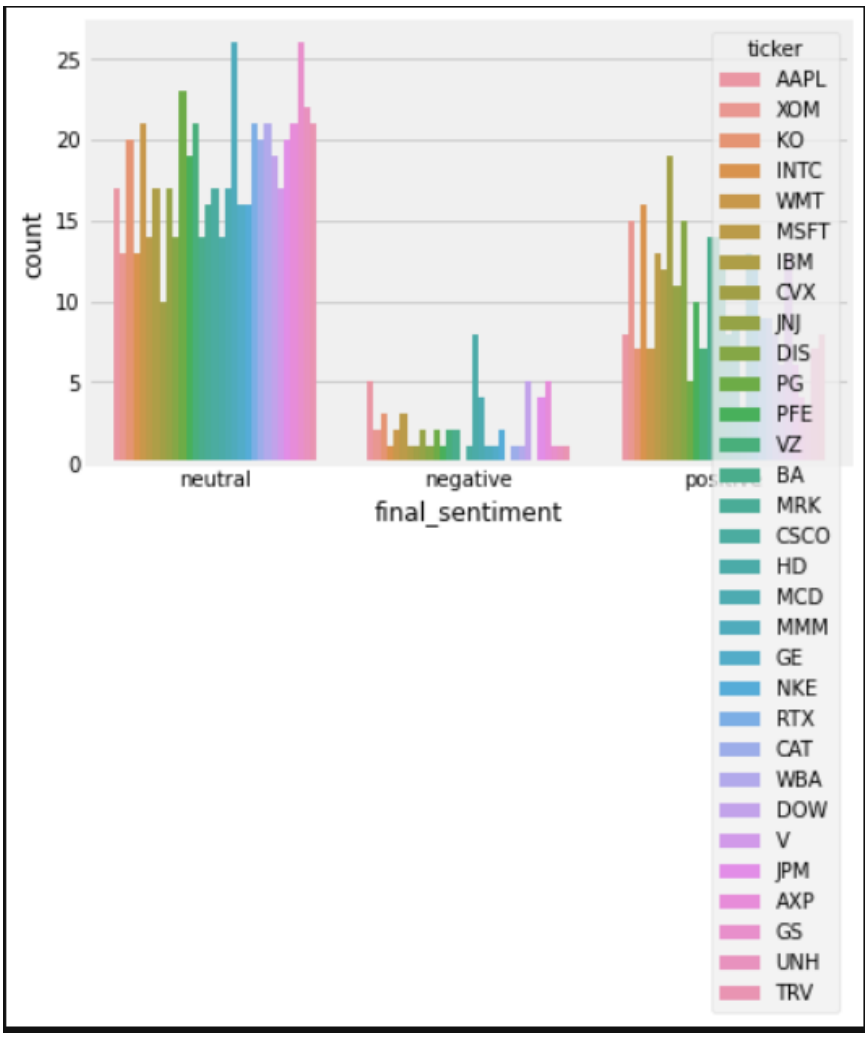
We can see that the distribution of negative, neutral and positive data is nearly the same for these set of tickers.

For headlines obtained from Google News API:



Most of the headlines were tagged as neutral, followed by positive by FinBERT.

For recent messages obtained from StockTwits JSON URL:



Most of the headlines were tagged as neutral, followed by positive by FinBERT.

Prediction methods and Results:

Algorithms	Description
NLTK Vader	VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.
FinBERT	A language model based on BERT, to tackle NLP tasks in financial domain. It has been fine tuned for this domain by training it on datasets like: TRC2-financial (a subset of Reuters' TRC2), Financial PhraseBank and FiQA Sentiment. BERT is in essence a language model that consists of a set of Transformer encoders stacked on top of each other. However, it defines the language modeling task differently from ELMo and AWD-LSTM. Instead of predicting the

	next word given previous ones, BERT "masks" a randomly selected 15% of all tokens. With a softmax layer over vocabulary on top of the last encoder layer the masked tokens are predicted. A second task BERT is trained on is "next sentence prediction". Given two sentences, the model predicts whether these two follow each other. The input sequence is represented with token and position embeddings. Two tokens denoted by [CLS] and [SEP] are added to the beginning and end of the sequence, respectively. For all classification tasks, including the next sentence prediction, [CLS] token is used. BERT has two versions: BERT-base, with 12 encoder layers, hidden size of 768, 12 multi-head attention heads and 110M parameters in total and BERT-large, with 24 encoder layers, hidden size of 1024, 16 multi-head attention heads and 340M parameters. Both models have been trained on BookCorpus and English Wikipedia, which have in total more than 3,500M words.

For testing the unlabeled FINVIZ headlines dataset, the VADER module from NLTK was used to perform this unsupervised learning task.

For testing the unlabeled Intrinio headlines dataset and Google News headlines dataset, the FinBERT model was used to perform these unsupervised learning tasks. The results returned more neutral and positive sentiment than negative. This is because, popular companies have been chosen which have a good upward stock price trend in recent times.

For testing the semi labeled StockTwits headlines dataset, the FinBERT model was used to perform this semi-supervised learning task.

The results from this semi-supervised learning task were as follows (where pre-tagged posts were predicted again with FinBERT whose output is negative, neutral or positive):

	precision	recall	f1-score	support
negative	0.36	0.18	0.24	22
neutral	0.00	0.00	0.00	0
positive	0.92	0.14	0.24	242
accuracy			0.14	264
macro avg	0.43	0.11	0.16	264
weighted avg	0.87	0.14	0.24	264

These poor metrics are obtained since the datasets with which FinBERT was trained on does not reflect emojis used for financial context. Furthermore, financial phrases that are normally used in headlines are not used by social media users' posts like Tweets or messages. Lastly, given the lack of context that some messages have, it may be prudent to add the traditional data to get a better understanding of the tweets or messages posted by users on social media. Due to these reasons, FinBERT does not perform well on the StockTwits data. Another reason behind the poor metrics obtained is class imbalance.

A better approach at this stage was to predict the untagged data and merge the tagged data with the predicted untagged data to get a more complete and accurate understanding of the dataset, to predict the market movement in the future.

Conclusion:

Stock price prediction is a challenging area of study that has been an important study of study for centuries. Using alternative data, we may improve insights obtained from traditional (quantitative) data, due to the time lag nature of that data, which is hardly present in alternative like news headlines and social media posts. The FinBERT variant of BERT transformer model gives us promising results on data obtained from headlines but needs further fine tuning to predict the data from social media websites like StockTwits.

Future Work:

The next step would be to pretrain existing FinBERT for social media financial jargon. Beyond that we can obtain more data from StockTwits users (with StockTwits API) with high follower counts (say over a 1000) to give weights to their opinions based on the follower counts. Additionally, for alternative data, we can obtain from Google News API to integrate news regarding the supply chain and integrate that sentiment to this study, as it is a crucial aspect of reducing time lag and generating alpha in the stock market. Finally, these alternative studies can be integrated with quantitative analyses on market data and company financials, to obtain a complete understanding of the different factors that influence stock price, and we can formulate an appropriate trading strategy from this understanding.

Acknowledgement:

I would like to thank Dr. Wang for his unwavering encouragement and readily available support during the trying aspects of this project.

References:

- https://www.intechopen.com/online-first/recent-advances-in-stock-market-prediction-using-text-mining-a-survey
- https://www.investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/
- https://en.wikipedia.org/wiki/Alternative_data_(finance)
- https://ieeexplore.ieee.org/document/9142175
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0111-6
- https://link.springer.com/article/10.1186/s40854-020-00205-1
- https://medium.com/arteos-ai/welcome-bert-state-of-the-art-language-model-for-nlp-by-google-e88d2381be6c
- https://arxiv.org/abs/1810.04805
- [1908.10063] FinBERT: Financial Sentiment Analysis with Pre-trained Language Models (arxiv.org)
- https://learn.datacamp.com/projects/611