Predicting Stock Prices Using News Headlines, Tweets and Search Trends

Betul Mescioglu
Department of Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, USA
betulmescioglu@lewisu.edu

Dr. Sam Abuomar
Department of Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, USA
oabuomar@lewisu.edu

***Abstract*— Numerous applications aim to predict future stock prices using historical data, but they often fail to promptly reflect the impacts of company-related events or CEO activities in stock prices. To gain insights into such events, individuals typically rely on news channels, search engines, and social media platforms. This paper investigates the potential enhancement of predictive models by integrating information from these sources with historical price data. Specifically, we focus on two companies, Tesla and HomeDepot, and collect relevant Tweets, Google news, and Google search trends spanning a three-year period. For Tesla, we also include its CEO Elon Musk as well to observe whether the controversies surrounding him has any impact on predicting Tesla's stock prices. We employ Linear Regression and a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to predict stock prices.**

Keywords— stock prediction, sentiment analysis, deep learning model

I. INTRODUCTION

There have been many studies related to stock market prediction based on historic prices and social media sentiments. For example, Quanzhi Li et al. employs the sentiment-oriented word embeddings obtained from tens of millions of StockTwits posts to detect domain-specific word polarity [1]. In Huihui et al., the authors develop a model that constructs the tweet node network to extract more meaningful information out of tweet source structure, on top of obtaining tweets' semantic features [2]. In Bos et al. the authors develop a model that automatically builds financial sentiment lexicons using two main approaches: weighted versions of Pointwise Mutual Information (PMI) and methods to account for negation. For the weighted PMI approach, they calculated PMI scores for word pairs based on their co-occurrence in a corpus of financial microblog messages from StockTwits. To address negation, the authors proposed two methods: the Negated Word approach and the Flip Sentiment approach. The Negated Word approach identified negation cues and modified the sentiment scores of words accordingly. The Flip Sentiment approach reversed the polarity of sentiment scores when negation cues were detected. The results showed that the developed model outperformed other sentiment lexicons [3]. In Gite et al., the authors perform something very similar to what this paper achieves to do; they combine historical price data with financial news headlines from Pulse, which aggregates financial news from various sources and finally utilize CNN-LSTM to predict future prices [4]. They tokenize the news headlines and use word embeddings in order to create word vectors. However, this method gave a 14% less accuracy than an LSTM applied to historical price data only.

In our study, we leverage the market's responsiveness to new information disseminated through influential platforms such as financial news and prominent users on Twitter such as Bloomberg and MarketWatch. When controversies arise concerning specific companies, individuals frequently resort to Google search for further investigation. The objective of our study is to examine the potential enhancement in predictive power of models when historical price data is enhanced by integrating sentiment analysis of news and tweets, as well as fluctuations in search trends related to a particular company.

To evaluate the influence of news, tweets, and search trends, we chose two companies: Tesla and HomeDepot. Tesla, being at the forefront of innovative technology and its CEO, Elon Musk, being involved in notable controversies, leads to significant media coverage. Our hypothesis is that this additional information could lead to more accurate price predictions for Tesla. On the other hand, HomeDepot, known for its stable leadership and comparatively lower engagement in news cycles compared to Tesla, serves as a benchmark for comparison. If our hypothesis is true, this additional information would not increase the predictive power of the models for HomeDepot.

To further distinguish Elon Musk's involvement in the changes of price, we run the same analysis on data obtained by excluding "Elon Musk" as a search keyword and excluding "Tesla" as a search keyword.

Specifically, we ran the analysis on:

- Historical data of Tesla
- Historical data, news, tweets, search trends related to Tesla, TSLA and Elon Musk (referred to as "Tesla + Elon Musk")

- Historical data, news, tweets, search trends related to only Tesla and TSLA (referred to as "Tesla Only"). "Elon Musk" is excluded.
- Historical data, news, tweets, search trends related to only Elon Musk (referred to as "Elon Musk Only"). "Tesla" and "TSLA" are excluded.
- Historical data of HomeDepot
- Historical data, news, tweets, search trends related to HomeDepot

II. Data Collection

Data were collected from four sources: historical price data from Yahoo Finance, Google News, Twitter and Google search trends spanning from May 29, 2020 to May 29, 2023 utilizing yahoo_fin, pygooglenews, snscrape.modules.twitter, and pytrends libraries of Python respectively.

A. Historical Price Data:

Historical price data obtained from Yahoo Finance includes open (opening price at the beginning of a day), high (highest traded price within a day), low (lowest traded price within a day), close (final traded price at the end of a day), adjusted close price (price accounting for any actions or events affecting the stock price) and volume (total number of shares or contracts traded during a day). The stock market is open on weekdays, excluding holidays, resulting in a collection of 737 data points for each company.

B. Google News:

Google news published within the mentioned dates are investigated and news containing the desired keywords such as Tesla, Elon Musk are collected [5]. Only headlines of these news are used.

C. Twitter:

Two prominent Twitter accounts, @markets (Bloomberg) and @marketwatch (MarketWatch), with followers of 1.5 million and 4.4 million, respectively, are selected to monitor company-related tweets [6][7]. Tweets published by these accounts during the mentioned time span are investigated, and tweets containing the specific keywords are collected.

D. Google Trends:

Google search trends for the specific keywords are collected for the given time span [8].

III. Data Processing

A. Historical Price Data:

Historical data contain numerical values only, therefore other than scaling the data, no further preprocessing is needed. Since the task is predicting future prices, two target features "twoweeks" and "month" are generated by taking rolling average of "close" feature for 10 days and 20 days respectively. In the analysis, "twoweeks" feature is predicted, however, "month" is kept to observe how correlations change over time.

B. Google News:

Google News gathers news from various respectable news sources. We only use headlines which were carefully crafted. Therefore, no cleaning is necessary.

C. Tweets:

Similar to news, Twitter accounts @markets and @marketwatch carefully craft their tweets. Other than removing escape sequences, there is no need for further cleaning.

D. Google Trends:

These are numerical data corresponding to how many times a certain keyword is searched during a week. If there were more than one search word such as Tesla and Elon Musk, we add all search words' trends. Although these values are provided on a weekly basis only, this does not pose a problem as we aggregate the rest of the data on a weekly basis as well.

E. Sentiment Analysis of Tweets and News:

To choose the optimal sentiment analysis tool, three sentiment analysis techniques were investigated; ChatGPT, Vader Sentiment Analyzer from nltk library of Python, and Twitter-roBERTa-base model [9]. A sample of 100 tweets were analyzed using these three techniques. In Table I, we can observe how these tweets are labeled by each model:

TABLE I. Sentiment Analysis Tools

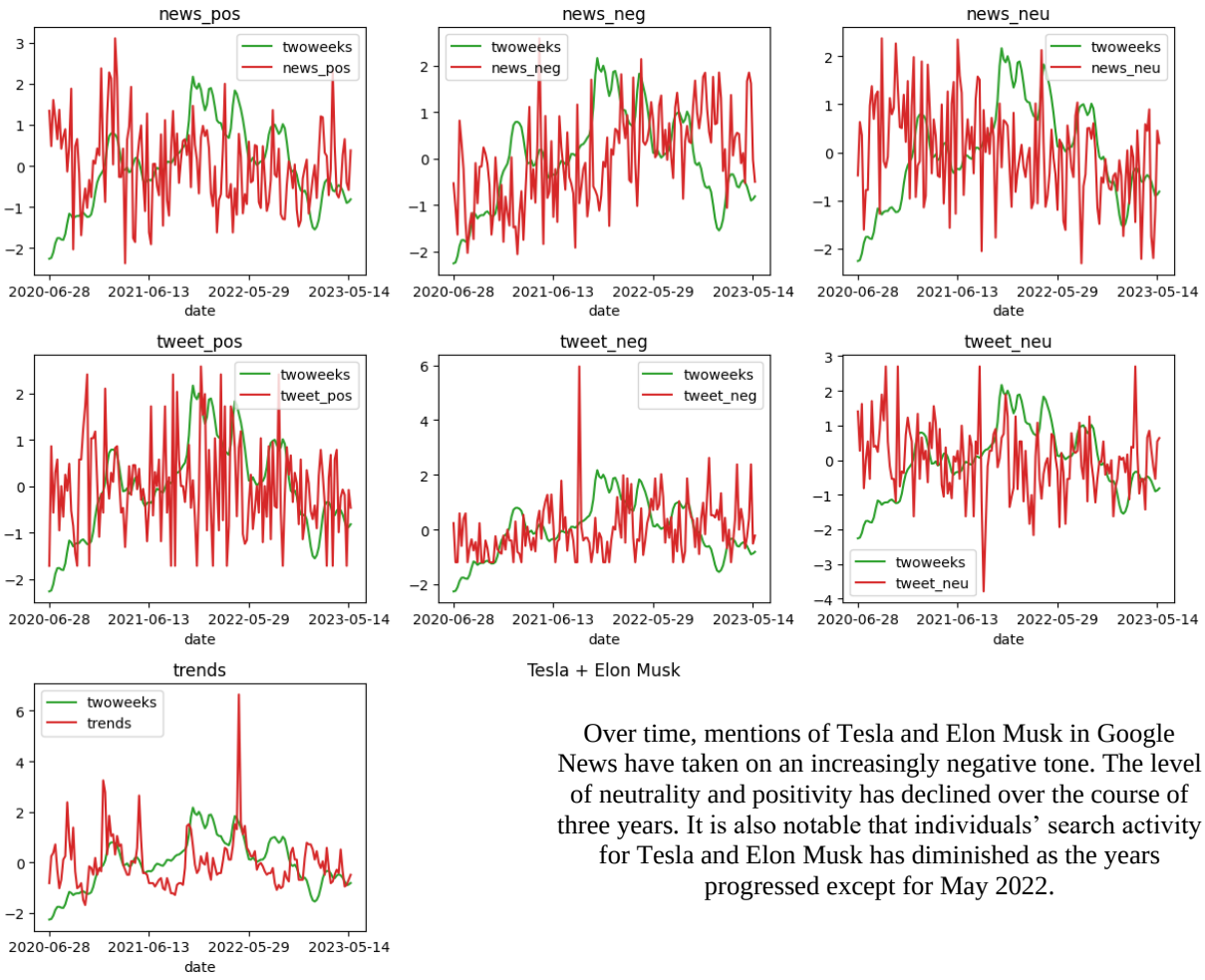
ChatGPT		
Neu.	***Pos.***	***Neg.***
69	16	15
Vader		
Neu.	***Pos.***	***Neg.***
50	28	22
roBERTa-base		
Neu.	***Pos.***	***Neg.***
59	24	17

We would like more polarity in our answers to improve the predictive power of the model. Vader Analyzer labels fewer tweets as neutral than ChatGPT and roBERTa-base. Additionally, roBERTa-base and ChatGPT put a rate limit on requests. Due to these reasons, we decide to employ the Vader Sentiment Analyzer in determining the sentiment of the news and tweets. The analyzer assigns a numerical value between -1 and 1 to represent the sentiment, with 1 indicating a highly positive sentiment, -1 indicating a highly negative sentiment, and 0 indicating neutrality. We divide this range into three equal sections and label the text as 'pos', 'neu', or 'neg' based on which section their sentiment value falls into. The data is then aggregated on a weekly basis, and the sentiment values for each group are counted and used as weights for that group. For example, if 50% of tweets exhibit a positive tone, 20% are neutral, and 30% are negative in a given week, the values for tweets_pos, tweets_neu, and tweets_neg would be assigned as 0.5, 0.2, and 0.3, respectively.

Finally, historical price data, trends, news and tweets sentiment data are brought together and scaled using Standard Scaler. 70% of the data is used for training set and the rest is allocated as the testing set.

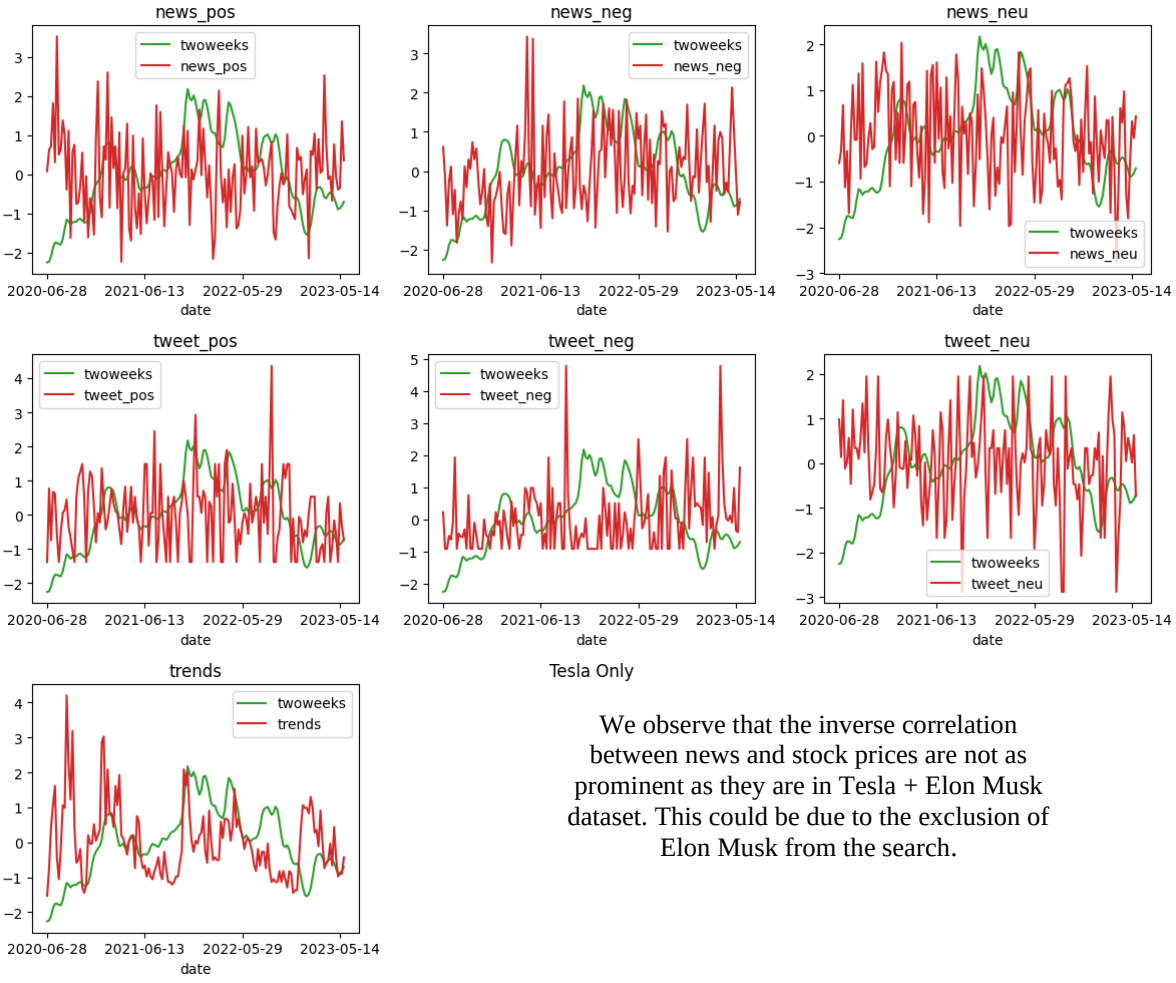
IV. Data Exploration

The following figures illustrate the changes in sentiment data and trends data in relation to stock data over the three-year period for all datasets:



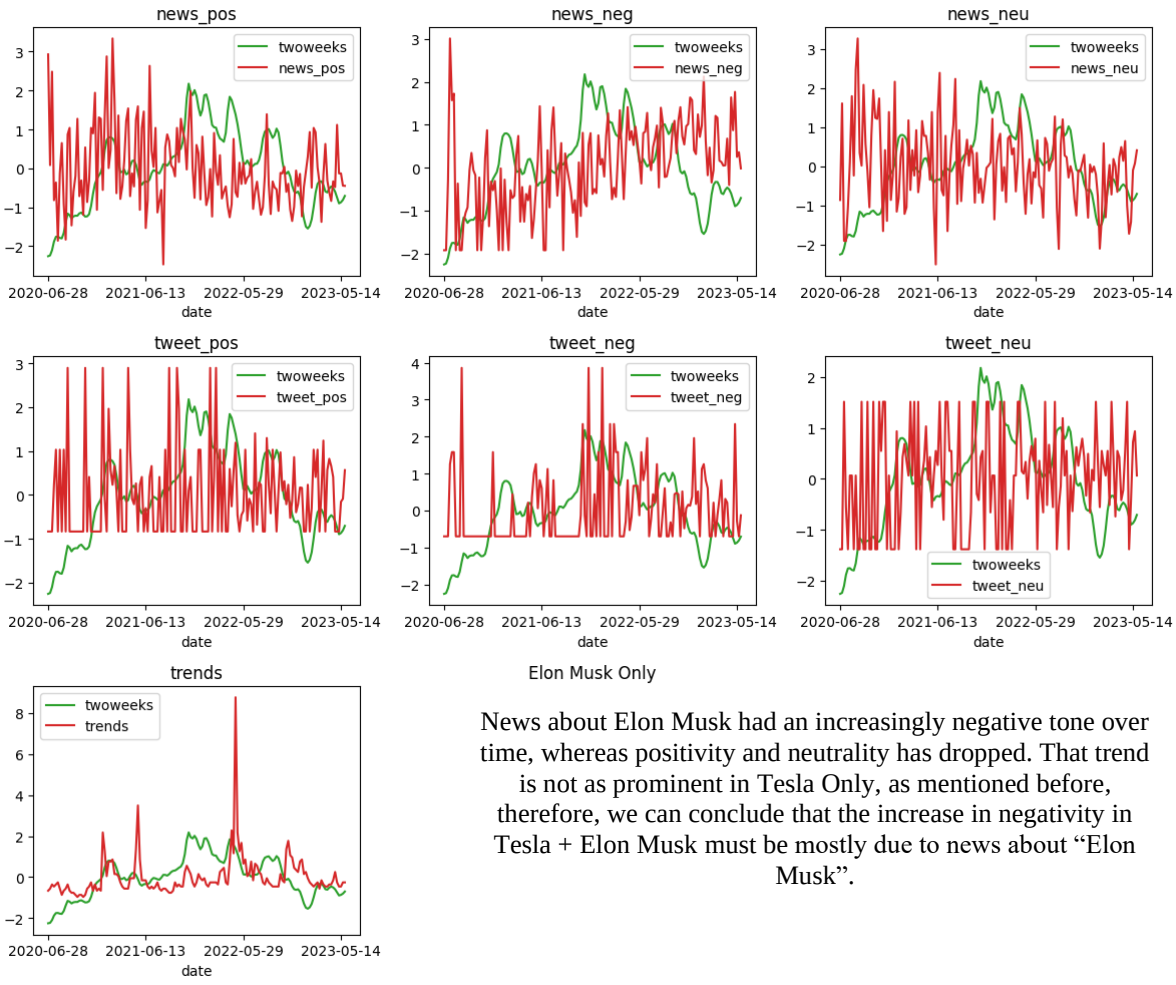
Over time, mentions of Tesla and Elon Musk in Google News have taken on an increasingly negative tone. The level of neutrality and positivity has declined over the course of three years. It is also notable that individuals' search activity for Tesla and Elon Musk has diminished as the years progressed except for May 2022.

Fig. 1. Tesla + Elon Musk sentiment and trends vs stock price



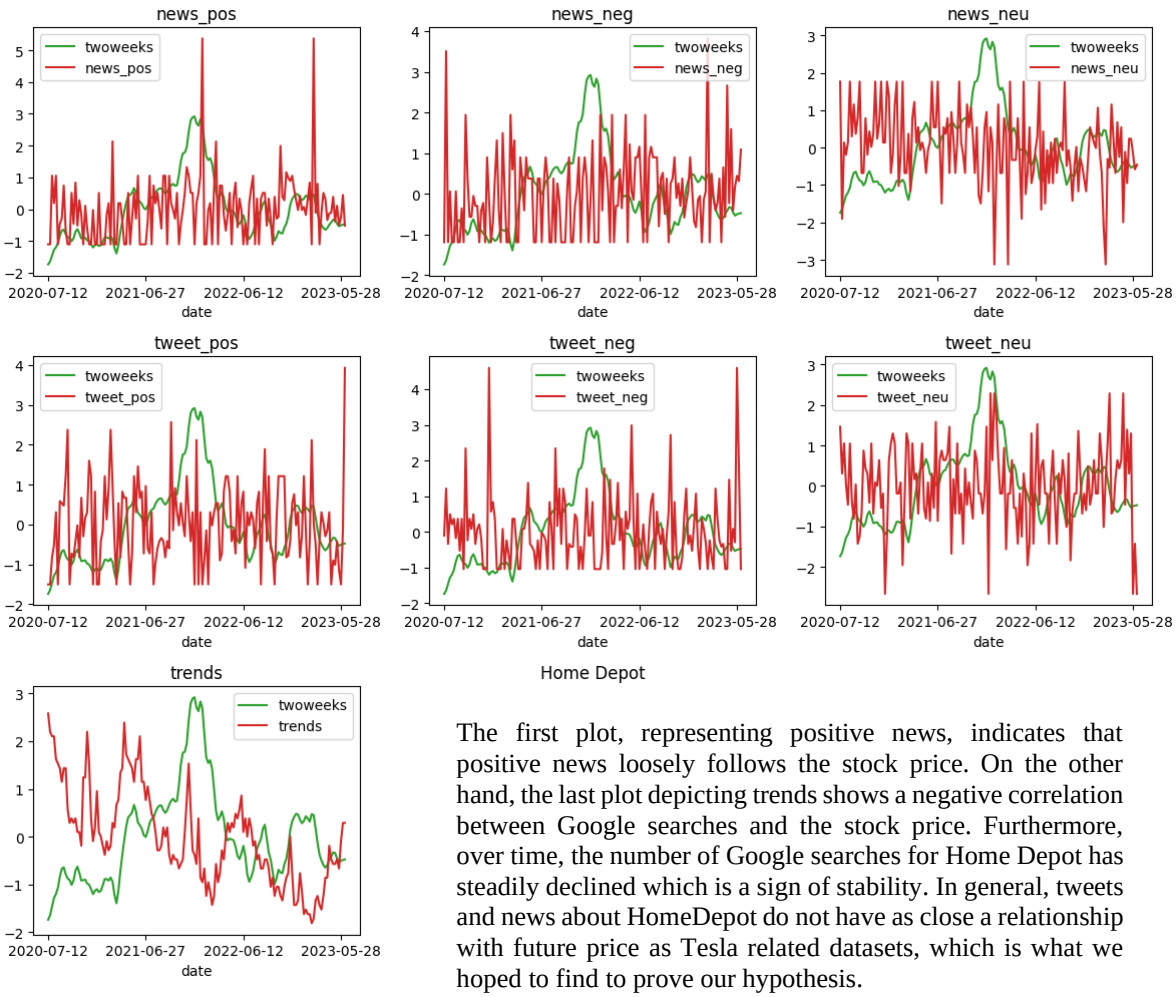
We observe that the inverse correlation between news and stock prices are not as prominent as they are in Tesla + Elon Musk dataset. This could be due to the exclusion of Elon Musk from the search.

Fig. 2. Tesla Only sentiment and trends vs stock price



News about Elon Musk had an increasingly negative tone over time, whereas positivity and neutrality has dropped. That trend is not as prominent in Tesla Only, as mentioned before, therefore, we can conclude that the increase in negativity in Tesla + Elon Musk must be mostly due to news about “Elon Musk”.

Fig. 3. Elon Musk Only sentiment and trends vs stock price



The first plot, representing positive news, indicates that positive news loosely follows the stock price. On the other hand, the last plot depicting trends shows a negative correlation between Google searches and the stock price. Furthermore, over time, the number of Google searches for Home Depot has steadily declined which is a sign of stability. In general, tweets and news about HomeDepot do not have as close a relationship with future price as Tesla related datasets, which is what we hoped to find to prove our hypothesis.

Fig. 4. Home Depot sentiment and trends vs stock price

V. FEATURE SELECTION THROUGH FEATURE CORRELATIONS AND FACTOR ANALYSIS

A. Analysis of Feature Correlations:

TABLE II. FEATURE CORRELATIONS WITH TARGET VARIABLES

	Tesla + Elon Musk		***Tesla Only***		***Elon Musk Only***		***HomeDepot***	
	Two Wk.	**Mon.**	**Two Wk.**	**Mon.**	**Two Wk.**	**Mon.**	**Two Wk.**	**Mon.**
open	0.986 <.0001	0.956 <.0001	0.986 <.0001	0.956 <.0001	0.986 <.0001	0.956 <.0001	0.983 <.0001	0.954 <.0001
high	0.985 <.0001	0.955 <.0001	0.985 <.0001	0.955 <.0001	0.985 <.0001	0.955 <.0001	0.982 <.0001	0.954 <.0001
low	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.978 <.0001	0.947 <.0001
close	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.978 <.0001	0.949 <.0001
adjcl.	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.981 <.0001	0.950 <.0001	0.970 <.0001	0.945 <.0001
vol.	-0.644 <.0001	-0.671 <.0001	-0.645 <.0001	-0.671 <.0001	-0.645 <.0001	-0.671 <.0001	0.018 0.8263	0.058 0.4785
news_pos	-0.053 0.5162	-0.078 0.3413	-0.097 0.2345	-0.119 0.1424	-0.008 0.9239	-0.026 0.7484	0.248 0.0020	0.229 0.0044
news_neg	0.088 0.2792	0.141 0.0832	0.065 0.4213	0.100 0.2195	0.044 0.5864	0.081 0.3175	-0.035 0.6681	-0.034 0.6805
news_neu	-0.044 0.5942	-0.075 0.3612	0.016 0.8477	0.003 0.9744	-0.032 0.6961	-0.048 0.5547	-0.160 0.0484	-0.146 0.0708
tweet_pos	0.112 0.1704	0.095 0.2464	0.143 0.0779	0.120 0.1385	0.085 0.2960	0.091 0.2612	0.006 0.9458	-0.003 0.9678
tweet_neg	0.016 0.8437	0.052 0.5218	-0.089 0.2722	-0.063 0.4361	0.073 0.3708	0.090 0.2693	-0.077 0.3453	-0.074 0.3622
tweet_neu	-0.120 0.1397	-0.137 0.0924	-0.044 0.5852	-0.047 0.5609	0.046 0.5739	0.059 0.4665	0.062 0.4430	0.068 0.4022
trends	0.198 0.0144	0.187 0.0214	0.037 0.6490	-0.001 0.9883	0.281 0.0004	0.313 <.0001	-0.266 0.0009	-0.311 <.0001

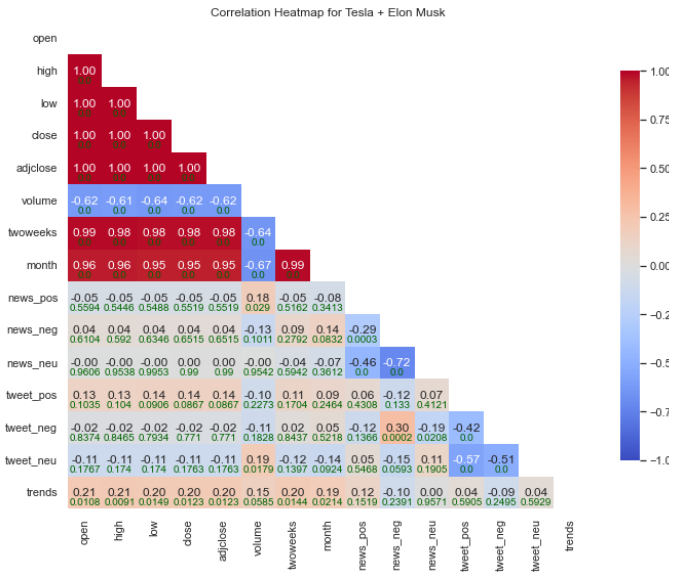


Fig. 5. All feature correlations for Tesla + Elon Musk

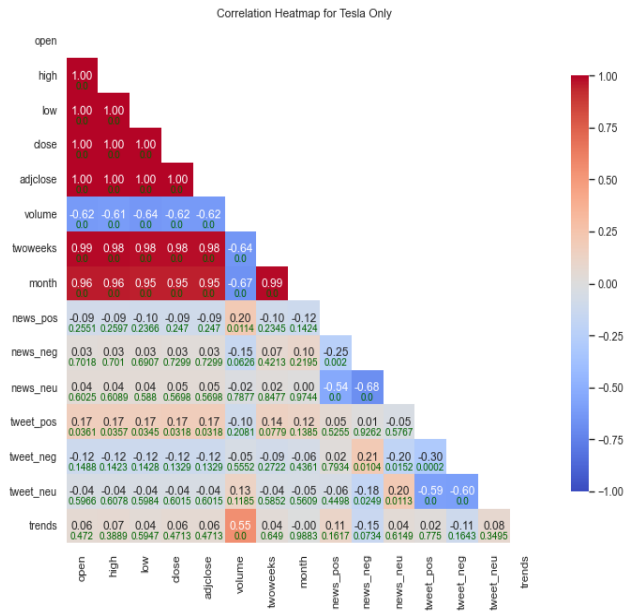


Fig. 6. All feature correlations for Tesla Only

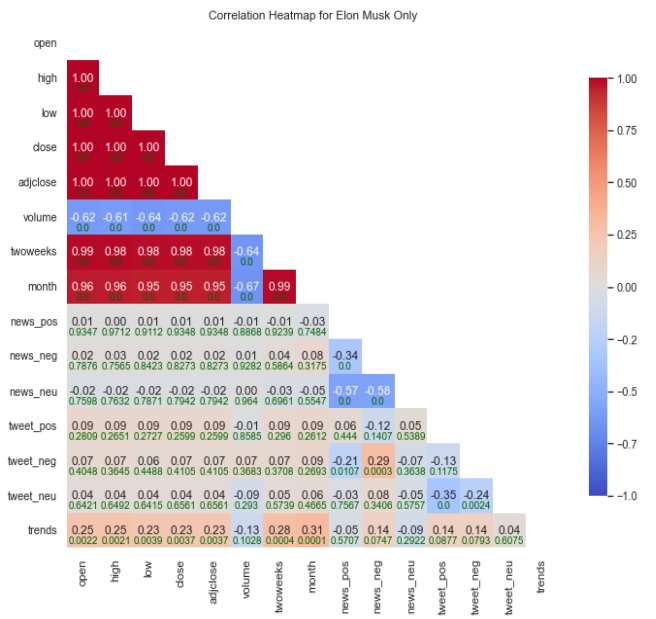


Fig. 7. All feature correlations for Elon Musk Only

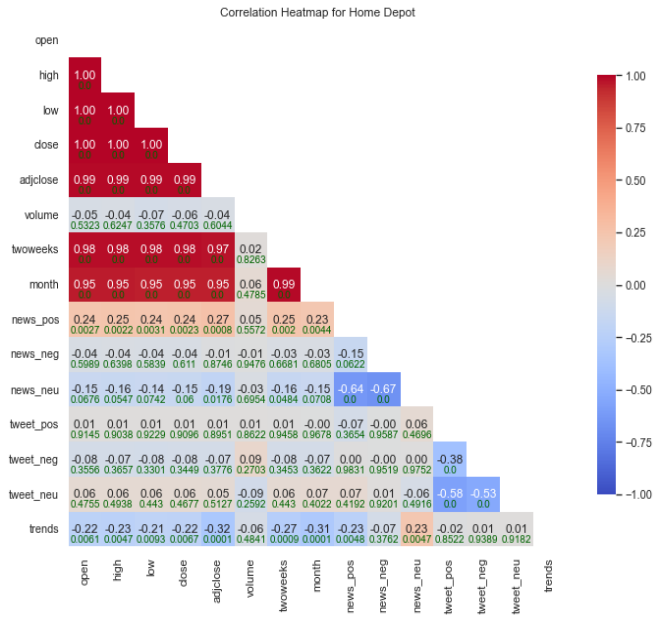


Fig. 8. All feature correlations for HomeDepot

From Table II and Figures 5-8, it can be observed that target values "twoweeks" and "month" are highly correlated with historical data, specifically "open" through "volume" for Tesla and/or Elon Musk related datasets. Interestingly, "volume" is negatively correlated with the price increase over two weeks and one month. For HomeDepot, on the other hand, there is no correlation between "volume" and future price; however, other historical data is highly correlated with future price.

In Figure 5, for the Tesla + Elon Musk dataset, among sentiment and trends data, the only statistically significant correlation with "twoweeks" is seen with the "trends" feature at 0.20. The same level of correlation is not observed in the Tesla Only dataset; however, it appears even stronger in the Elon Musk Only dataset at 0.28. There is a statistically significant negative correlation between "trends" and "twoweeks" in the HomeDepot dataset. This finding is crucial in proving our hypothesis, as when there are ominous signs, people tend to research the company, and panic can cause prices to fall. For a stable company, it is expected that higher search volume would result in lower prices in two weeks, which is indeed observed with HomeDepot. In contrast, Tesla's price actually increases in two weeks when people start conducting research, although this is not due to Tesla searches. Searches about Elon Musk are the driving factor behind the price increase, as excluding Tesla searches strengthens the correlation. Elon Musk's personal impact on the price increase would certainly contribute to enhancing the predictive power of the models.

Another indicator of stability is that an increase in positive news about a company is expected to lead to a price increase in two weeks. In Figure 8, this is observed for HomeDepot. Although the correlation is weak at 0.25, there is still a statistically significant positive correlation between positive news and price over two weeks. However, for Tesla-related datasets, there is minimal correlation, and this correlation indicates an inverse relationship. This inverse relationship is also evident in Figures 1 and 3. However, this does not hold true for tweets. Nevertheless, these findings related to Tesla datasets are not statistically significant; hence, we cannot base our judgments on these findings.

When selecting features for the Linear Regression model, we aim to avoid multicollinearity, which is evident among historical data variables. The variable "open" is correlated with "twoweeks" at a higher value across all datasets; therefore, "open" is chosen to represent all price columns. Additionally, variables that are correlated with "twoweeks" at a low p-value should be chosen. Therefore, for the Tesla + Elon Musk dataset, variables "open," "volume," and "trends" are selected; for the Tesla Only dataset, "open" and "volume" are selected; for the Elon Musk Only dataset, "open," "volume," and "trends" are chosen; and finally, for the HomeDepot dataset, "open," "news_pos," and "trends" variables are selected.

One would expect confidence to increase and errors to decrease when including only statistically significant variables, compared to including all variables. However, as shown in Table III, this is not the case. The reason for this could be the presence of non-linear relationships between the independent variables and the target variable. Correlation captures only the linear relationship. By excluding variables with low correlation, valuable information provided by these variables, which have a non-linear relationship with the target variable, is also omitted. For this reason, including all variables may be a better option.

TABLE III. LINEAR REGRESSION RESULTS WITH ALL VARIABLES AND SELECTED VARIABLES

	Tesla + Elon Musk		***Tesla Only***		***Elon Musk Only***		***Home Depot***	
	Conf.	**MSE**	**Conf.**	**MSE**	**Conf.**	**MSE**	**Conf.**	**MSE**
All Hist.	0.97	0.015	0.97	0.016	0.97	0.016	0.84	0.031
All Hist.+ Sent. + Trends	0.97	0.0146	0.97	0.015	0.97	0.0156	0.83	0.032
Selected Variables	0.961	0.021	0.96	0.0215	0.96	0.0199	0.82	0.034

B. Factor Analysis:

The Kaiser-Meyer-Olkin (KMO) measure of the datasets Tesla + Elon Musk, Tesla Only, Elon Musk Only, and HomeDepot are 0.62, 0.62, 0.63, and 0.71, respectively. While the datasets are not ideally suited for factor analysis, we still proceed with exploring the factors. The loadings and variances for all datasets are presented in the following tables:

TABLE IV. TESLA + ELON MUSK LOADINGS AND VARIANCES

	Tesla + Elon Musk factors:				
Factors:	0	1	2	3	4
low	0.995	0.004	0.04	-0.002	0.054
open	0.994	0.003	0.03	0.004	0.063
close	0.993	0.007	0.04	-0.003	0.061
adjclose	0.993	0.007	0.04	-0.003	0.061
high	0.993	0.003	0.03	0.006	0.068
volume	-0.69	0.09	-0.12	-0.06	0.38
news_neu	0.01	0.96	-0.02	-0.036	-0.26
news_neg	0.067	-0.82	-0.07	0.22	-0.31
tweet_pos	0.103	0.084	0.93	-0.29	0.02
tweet_neu	-0.08	0.082	-0.8	-0.6	0.03
tweet_neg	-0.02	-0.18	-0.09	0.95	-0.057
news_pos	-0.1	-0.27	0.11	-0.24	0.76
trends	0.192	0.197	-0.08	0.117	0.692
Variance	***5.48***	***1.78***	***1.55***	***1.47***	***1.38***
Propor. Var	***0.42***	***0.16***	***0.12***	***0.12***	***0.08***
Cumul. Var	***0.42***	***0.58***	***0.70***	***0.82***	***0.90***

The first factor carries information mostly related to historical data. The second factor carries mostly negative and neutral news. The third factor holds mostly positive and neutral tweets. The fourth factor holds mostly negative tweets. The fifth factor holds mostly positive news and trends. 90% of the variance of data is captured by these five factors.

TABLE V. TESLA ONLY LOADINGS AND VARIANCES

	Tesla Only factors:					
Factors:	0	1	2	3	4	5
high	0.997	0.002	-0.028	0.046	-0.01	-0.025
open	0.997	0.0007	-0.027	0.046	-0.024	-0.025
close	0.996	-0.002	-0.03	0.05	-0.023	-0.026
adjclose	0.996	-0.002	-0.03	0.05	-0.023	-0.026
low	0.996	0.0003	-0.03	0.05	-0.04	-0.027
news_neg	0.025	0.966	0.10	0.02	-0.082	-0.22
news_neu	0.029	-0.82	-0.098	-0.042	0.0096	-0.555
tweet_neg	-0.086	0.115	0.98	-0.15	-0.057	0.016
tweet_neu	-0.029	-0.1	-0.71	-0.69	0.047	-0.038
tweet_pos	0.122	0.006	-0.14	0.98	0.002	0.029
trends	0.091	-0.057	-0.041	0.004	0.97	0.035
volume	-0.62	-0.05	-0.07	-0.042	0.7	0.11
news_pos	-0.067	-0.023	0.013	0.033	0.08	0.99
Variance	***5.4***	***1.6***	***1.5***	***1.5***	***1.4***	***1.4***
Propor. Var	***0.42***	***0.17***	***0.12***	***0.11***	***0.09***	***0.08***
Cumul. Var	***0.42***	***0.59***	***0.71***	***0.82***	***0.91***	***0.99***

The first factor carries information mostly related to historical price data. The second factor carries mostly negative and neutral news. The third factor holds mostly negative and neutral tweets. The fourth factor holds mostly positive tweets. The fifth factor holds mostly volume and trends data. Finally, the sixth factor holds mostly positive news. 99% of the variance of data is captured by these six factors.

TABLE VI. ELON MUSK ONLY LOADINGS AND VARIANCES

	Elon Musk Only factors:			
Factors:	0	1	2	3
low	0.99	0.065	-0.004	-0.034
open	0.99	0.075	-0.004	-0.035
adjclose	0.99	0.07	-0.002	-0.038
close	0.99	0.07	-0.002	-0.04
high	0.99	0.081	-0.002	-0.038
volume	-0.71	0.074	0.014	-0.1
news_neg	-0.03	0.84	-0.183	0.27
tweet_neg	0.004	0.71	0.18	-0.17
trends	0.25	0.35	-0.04	-0.14
news_neu	0.007	-0.4	0.9	-0.1
news_pos	0.02	-0.38	-0.85	-0.14
tweet_neu	0.088	-0.168	-0.046	0.8
tweet_pos	0.086	-0.099	-0.061	-0.76
Variance	**5.48**	**1.72**	**1.61**	**1.43**
Propor. Var	***0.42***	***0.135***	***0.12***	***0.11***
Cumul. Var	***0.42***	***0.56***	***0.68***	***0.79***

First factor captures historical data. The second factor carries mostly negative news and negative tweets. The third factor holds positive and neutral news. The fourth factor holds neutral and positive tweets. Trends data is not incorporated in any of the factors. Only 79% of the variance of data is captured by these four factors.

TABLE VII. HOMEDEPOT LOADINGS AND VARIANCES

	HomeDepot factors:				
Factors:	0	1	2	3	4
low	0.99	0.007	-0.01	0.07	-0.037
close	0.99	0.01	-0.009	0.08	-0.033
open	0.99	0.006	-0.008	0.08	-0.03
high	0.99	0.01	-0.006	0.09	-0.03
adjclose	0.98	0.04	-0.00044	0.14	-0.03
news_neg	-0.043	0.97	0.029	-0.14	0.018
news_neu	-0.093	-0.8	0.08	-0.55	0.039
tweet_pos	0.008	-0.014	0.97	0.008	-0.25
tweet_neu	0.032	0.007	-0.76	-0.035	-0.64
news_pos	0.17	0.06	-0.14	0.87	-0.07
volume	-0.11	-0.14	0.09	0.39	0.19
trends	-0.22	-0.1	-0.06	-0.5	0.02
tweet_neg	-0.045	0.007	-0.14	0.03	0.98
Variance	**5.02**	**1.6**	**1.56**	**1.53**	**1.48**
Propor. Var	***0.4***	***0.14***	***0.13***	***0.11***	***0.08***
Cumul. Var	***0.4***	***0.54***	***0.67***	***0.78***	***0.86***

The first factor carries information mostly related to historical price data. The second factor carries mostly negative and neutral news. The third factor holds mostly positive tweets and neutral tweets. The fourth factor holds mostly positive news. Finally, the fourth factor represents negative tweets. Volume and trends data are not represented by any of the factors at the threshold of 0.6. 86% of the variance of data is captured by these six factors.

Interestingly, in certain cases, certain variables remain unaccounted for by any of the factors. For instance, when focusing solely on the Elon Musk dataset, even though trend data exhibits a higher correlation with the target than any other variable, it does not find representation among any of the factors. Furthermore, only 79% of the variance is explained by the factors in this specific dataset.

Another crucial point is that the primary objective of sentiment analysis is to categorize the meaning of text into distinct sentiment levels such as positive, neutral, or negative. This categorization aids the model in grouping similar information together, thereby enhancing its predictive capability. However, during factor analysis, these sentiment levels are frequently amalgamated. This blending is particularly evident with neutral tweets and neutral news, as they become amalgamated with either positive or negative tweets and news, leading to a loss of their individual significance.

Given these factors, it is reasonable to anticipate that models generated using these factors might yield weaker results. Indeed, the Linear Regression model developed with these factors demonstrates a higher error rate with notably reduced confidence when compared to a model employing all available

variables for the Tesla + Elon Musk dataset. Table VIII effectively portrays these outcomes:

TABLE VIII. LINEAR REGRESSION RESULTS USING ALL VARIABLES AND FACTORS

	All Variables		*Factors*	
	Conf.	**MSE**	**Conf.**	**MSE**
Historical Data	0.97	0.015	0.915	0.0457
Historical + Sentiment + Trends Data	0.97	0.0146	0.923	0.0417

The outcomes from both the exploration of feature correlations and the factor analysis unequivocally suggest that employing all variables is the more effective approach for capturing the maximum variance in the data and attaining lower error rates.

VI. METHODOLOGY

As prediction methods, Linear Regression and a combination of CNN+LSTM are used on only historical price datasets and datasets including historical and sentiment and trends data of each company. In the hopes of increasing predictive power, K-Means clustering method is employed to obtain the clusters of data points. This information is appended to datasets and the same analysis is performed once again. The goal is to predict stock prices for the next two weeks.
The following summarizes these steps:

- Linear Regression on historical stock price of Tesla
- Linear Regression on historical stock price of Home Depot.
- Linear Regression on "Tesla + Elon Musk" historical + sentiment + trends
- Linear Regression on "Tesla Only" historical + sentiment + trends
- Linear Regression on "Elon Musk Only" historical + sentiment + trends
- Linear Regression on "Home Depot" historical + sentiment + trends
- CNN + LSTM on historical stock price of Tesla
- CNN + LSTM on historical stock price of Home Depot
- CNN + LSTM on "Tesla + Elon Musk" historical + sentiment + trends
- CNN + LSTM on "Tesla Only" historical + sentiment + trends
- CNN +LSTM on "Elon Musk Only" historical + sentiment + trends
- CNN + LSTM on "Home Depot" historical + sentiment + trends
- K-Means on all four datasets to determine the clusters
- Perform the same analysis mentioned above on the datasets including the clusters.

A. Linear Regression:

Linear Regression is commonly used in predicting stock prices because it provides a simple and interpretable model that can capture linear relationships between predictor variables and the target variable. Linear Regression serves as a baseline model for comparing the results of CNN+LSTM in our analysis. Linear Regression results along with corresponding mean squared error (MSE) and confidence measures are presented in the "Results" section of this paper.

B. CNN + LSTM:

The CNN + LSTM with the following sequential structure is used in making predictions using the historical data (open, high, low, close, adjclose, volume) [10]. With addition of new features, input shape was increased to 13.

a. CNN Layers:

Convolution uses kernels to slide over the input data, performing element-wise multiplication and summing the results to produce a feature map. This process helps the model capture the relationships within the input data. Max-pooling is used to downsample the feature maps, reducing their dimensions while retaining the most relevant information. This process is repeated three times. Finally, the shape of the output is reduced to a 1D vector.

b. LSTM Layers:

LSTM is a type of recurrent neural network architecture. It is specifically designed to handle sequence data, such as time series, where preserving long-term dependencies is crucial. They have a unique memory cell that allows them to retain and update information for long durations, enabling them to learn patterns and relationships in the time series data. In our structure we have two layers of LSTMs. Dropout is a regularization technique that randomly sets a fraction of input units to 0 during training to prevent overfitting.

Finally, a fully connected layer with a linear activation function is used to produce the final output. The model is compiled with the Adam optimizer, MSE loss function, and MSE and mean absolute error (MAE) as evaluation metrics. In the results section we present only MSE metric.

The architecture of the CNN + LSTM model can be seen in Figure 9:

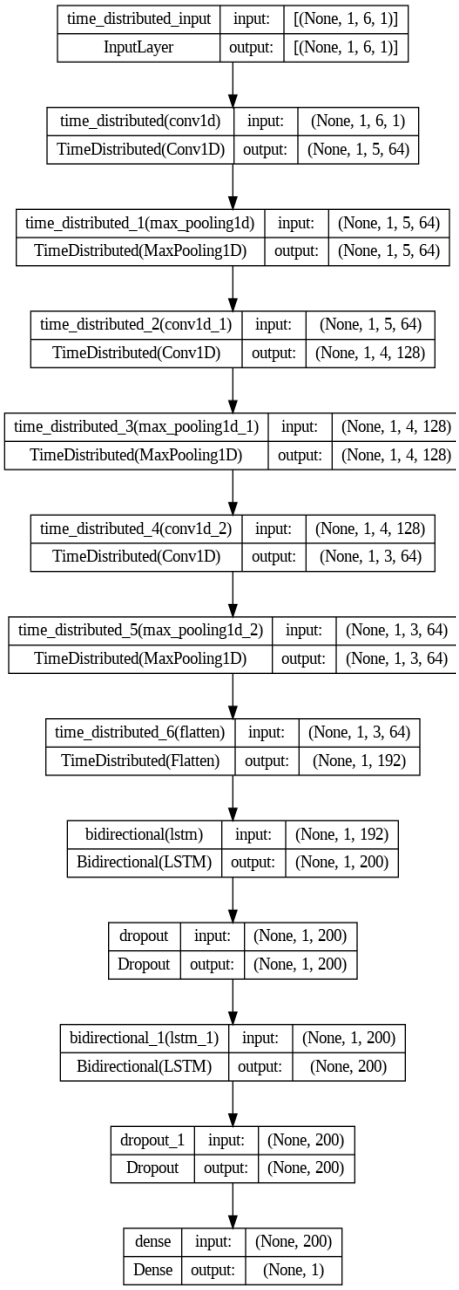


Fig. 9. CNN + LSTM architecture

C. *KMeans Clustering:*

In an effort to learn how historical data and sentiment and trends data integration occurs, we employ KMeans clustering model. We reduce the dimension of the dataset to three features using PCA and search k values in the range of 1 to 10. KElbowVisualizer determined the best number of clusters to be 5 for Tesla and Elon Musk dataset which is presented in the following visual:

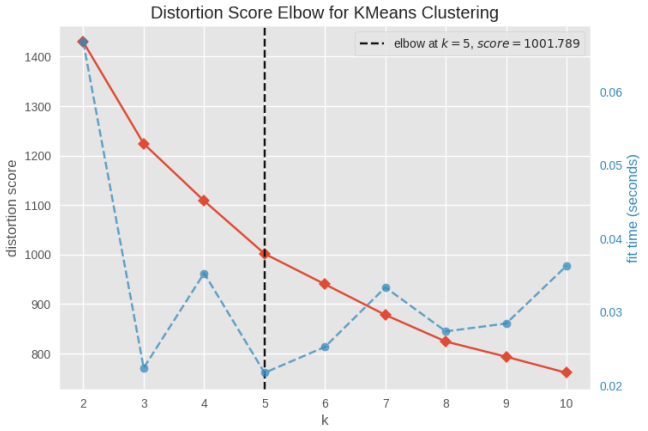


Fig. 10. Finding number of clusters for Tesla + Elon Musk

All datasets have five clusters except for Elon Musk Only which has four clusters. The following figure shows the visualizer result for Elon Musk Only:

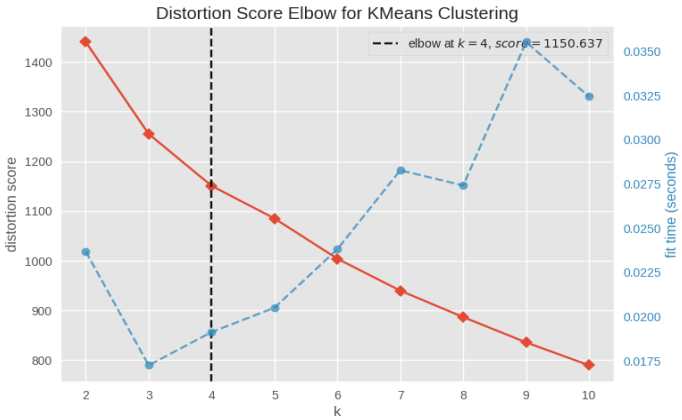


Fig. 11. Finding number of clusters for Elon Musk Only

VII. Results and Discussion:

A. *Linear Regression Results:*

TABLE IX. Linear Regression Results

	Tesla + Elon Musk		***Tesla Only***		***Elon Musk Only***		***Home Depot***	
	Conf.	**MSE**	**Conf.**	**MSE**	**Conf.**	**MSE**	**Conf.**	**MSE**
Historical Data	0.97	0.015	0.97	0.016	0.97	0.016	0.84	0.031
Historical + Sentiment + Trends Data	0.97	0.0146	0.97	0.015	0.97	0.0156	0.83	0.032
Clustered Historical Data	0.98	0.021	0.97	0.021	0.97	0.02	0.95	0.0312
Clustered Historical + Sentiment+Trends Data	0.97	0.023	0.97	0.019	0.97	0.02	0.95	0.0303

From Table IX, we observe that adding sentiment and trends data did not have much impact on the prediction error. Adding cluster data decreased the performance of the model for all datasets except for HomeDepot. The lowest error is observed when Linear Regression is applied to Tesla + Elon Musk dataset (Historical + sentiment + trends) without clustering. The following plots depict the results:

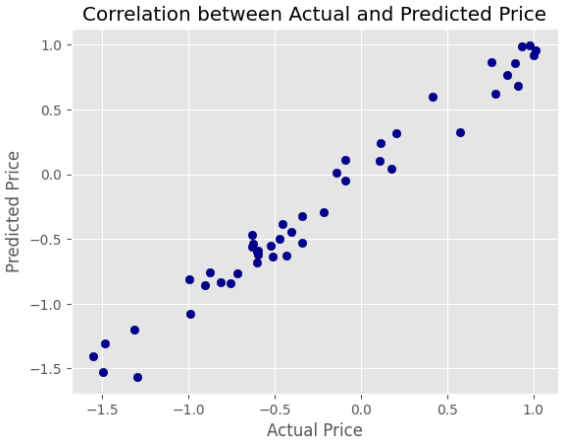


Fig. 12. Linear Regression results for Tesla + Elon Musk

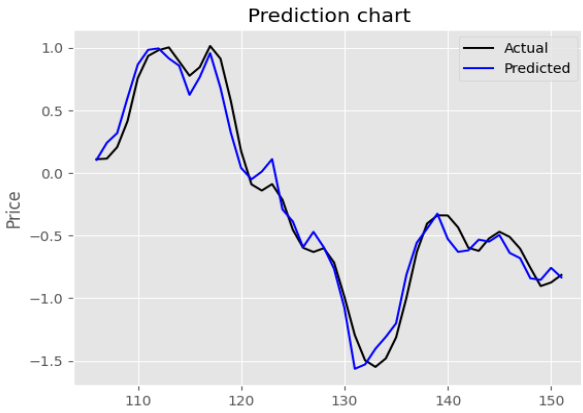


Fig. 13. Linear Regression test set predictions

From the prediction chart, we can see that the predicted values closely align with the observed values. However, there is still potential for further improvement.

B. *CNN + LSTM Results:*

TABLE X. CNN + LSTM RESULTS

	Tesla + Elon Musk		*Tesla Only*		*Elon Musk Only*		*Home Depot*	
	R2 Score	**MSE**	**R2 Score**	**MSE**	**R2 Score**	**MSE**	**R2 Score**	**MSE**
Historical Data	0.96	0.045	0.97	0.028	0.97	0.038	0.98	0.023
Historical + Sentiment + Trends Data	0.99	0.014	0.95	0.048	0.95	0.033	0.98	0.036
Clustered Historical Data	0.94	0.048	0.90	0.11	0.90	0.074	0.96	0.062
Clustered Historical + Sentiment+Trends Data	0.95	0.036	0.92	0.094	0.89	0.096	0.94	0.070

From Table X, we observe that CNN + LSTM outperforms Linear Regression in capturing the changes more effectively. Excluding Elon Musk from the analysis leads to an increase in the error of CNN + LSTM when incorporating sentiment and trends data. However, when both Elon Musk and Tesla are included in the search process, the error decreases by 69%. When only Elon Musk is included in the search process, the error remains approximately the same. For Home Depot, an increase in error is observed when sentiment and trends data are included. The addition of clusters negatively impacts the performance in all scenarios.

Figure 14 and Figure 15 illustrate the difference in precision between the predictions of the CNN + LSTM model when applied to historical data only and when applied to historical data along with sentiment and trends data respectively for the Tesla + Elon Musk dataset. We observe that prediction values are much more aligned with the actual values when the sentiment and trends data is included reflecting our finding of lower error.

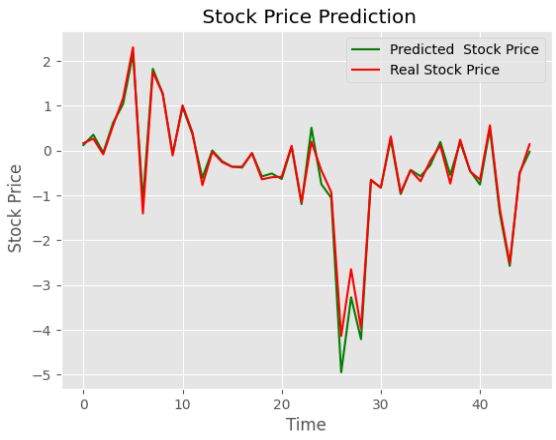


Fig. 14. CNN + LSTM results for Tesla historical price data

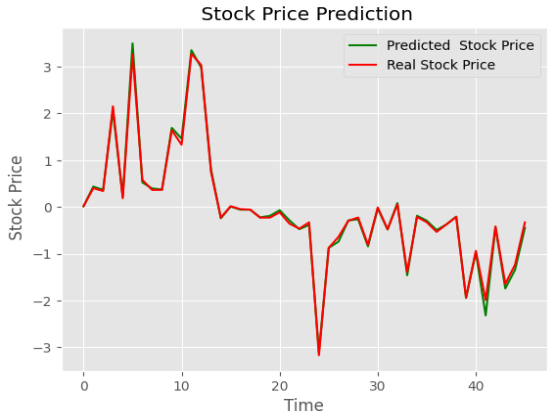


Fig. 15. CNN + LSTM results for Tesla historical price + sentiment + trends data

VIII. CONCLUSION

In this paper, we investigated whether the inclusion of news, tweets, and search trends pertaining to a company could enhance the predictive capability of a model for stock price prediction. Our study focused on two specific companies: Tesla and Home Depot. Tesla, renowned for pioneering technology, is associated with CEO Elon Musk, who has been involved in various controversies. In contrast, Home Depot exhibits more stable growth and is led in a comparatively steady manner.

Our hypothesis posited that the prominence of a company's name and its CEO in the news cycle would have a direct impact on stock prices. Consequently, incorporating machine learning algorithms with sentiment analysis and trend data would yield superior stock price predictions compared to relying solely on historical price, if the company is mentioned in the news and social media frequently. Our findings affirm the validity of this hypothesis for both Tesla and Home Depot.

The application of CNN + LSTM models supports our assertion that combining sentiment analysis and trend data with historical price data improves the algorithm's accuracy in predicting Tesla's stock prices. However, incorporating sentiment and trend data for Home Depot had an adverse effect on algorithm performance. Consequently, we conclude that this methodology is more suitable for companies that garner frequent attention in news circles, rather than for more stable enterprises.

REFERENCES

[1] Quanzhi Li and Sameena Shah. 2017. Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 301–310, Vancouver, Canada. Association for Computational Linguistics

[2] Huihui, N., Wang, S., & Cheng, P. (2021). A hybrid approach for stock trend prediction based on tweets embedding and historical prices. World Wide Web, 24(3), 849-868. doi:https://doi.org/10.1007/s11280-021-00880-9J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Bos, T. (Thomas), & Frasincar, F. (2021). Automatically Building Financial Sentiment Lexicons While Accounting for Negation. Cognitive Computation. doi:10.1007/s12559-021-09833-w

[4] Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N. Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Comput Sci. 2021 Jan 28;7:e340. doi: 10.7717/peerj-cs.340. PMID: 33816991; PMCID: PMC7924447.

[5] Google News (n. d.). Retrieved from https://news.google.com

[6] Twitter. (n. d.). markets [@markets]. Retrieved from https://twitter.com/markets

[7] Twitter. (n.d.). MarketWatch [@MarketWatch]. Retrieved from https://twitter.com/MarketWatch

[8] Google Trends (n. d.). Retrieved from https://trends.google.com/trends

[9] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

[10] Aadhitya, A. (2022). Stock Market prediction using CNN-LSTM. Retrieved from Stock Market prediction using CNN-LSTM | Kaggle