Green Investing using Alternative Data Sources

Kamila Zaman
New York University
kz2137@nyu.edu

Jorge Roldan-Roldan
New York University
jlr9718@nyu.edu

Antony Sunwoo
New York University
as10506@nyu.edu

ABSTRACT

In this work we explore the use of Twitter feeds and news articles as alternative data sources to predict the closing stock prices of leading green companies such as First Solar, Siemens, Plug Power and others. We show our methods to acquire data from alternative sources such as Twitter and news articles and perform sentiment analysis using NLTK over time. Using the time series obtained, we perform lag correlation and Granger causality, and TF-IDF analysis to investigate whether information gained from the alternative data sources and the stock market prices of leading green energy companies are in any way correlated.

1 INTRODUCTION

Data driven decision making is key to success. Living in this age of Big Data with numerous cutting edge technologies we have great opportunities at our disposal to make use of these resources to drive this data driven decision making approach into a reality and a culture for the current and upcoming industries. The rising global concern regarding climate change and the catastrophes it most probably might lead up to has given rise to several new companies being formed, organizations establishing new short-term/long-term policies and agreements;signed and presented to manage this issue with increased importance and priority. This mounting awareness is the cause for establishment of many new companies jumping into this industry along with the older ones improving, revamping and aligning their goals in agreement with climate change concerns as well. This gives us the perfectly timed opportunity to dive into putting forth an idea and methodology to explore the predictive power of the data for investment opportunities in this upcoming industry of green energy companies based on understanding at greater depths driven with creativity to analyze, comprehend and formulate an approach applicable to see forthcoming stock-trend's for individual companies and as whole groups to facilitate investors before hand to maximize developments and profits to make the companies with a good cause to grow further.

2 LITERATURE REVIEW

The usage of alternative data sources such as Twitter feeds, news articles and search queries for predicting financial markets has been increasingly receiving attention in industry and academia. The authors in [3] have used used Twitter feeds to investigate the correlation between the public mood and the value of the Dow Jones Industrial average. They used Opinion Finder (OF) and Google-Profile of Mood States (GPOMS) to track the sentiment mood collective mood. The authors also used a Granger causality analysis and Self-Organizing Fuzzy Neural Network to determine whether the collective mood is indeed correlated to the values of the DJIA, and successfully found an accuracy of 87.6 percent in predicting the daily behavior of this index.

Authors in [2] have used news articles besides Twitter feeds as an additional alternative data source. The authors used Opinion Finder (OF) and Stanford Natural Language Processing Programming Interface (SNLP) to do the sentiment analysis of these sources. Furthermore, they adopted a Time-based Term Frequency and Inverse Document Frequency (TF-IDF) to normalize the frequency of terms and used it to cluster terms using clustering algorithms. In this work the authors also did correlation analysis and Granger causality as well as other linear regression models to study the correlation between the collective mood from these data sources and the stock market behavior of various entities in Australia.

3 BUSINESS UNDERSTANDING

3.1 Background

It is now evident, as it has been for a long time that we are facing a climate crisis and there is not question about the urgent need to take action on this crisis. The Intergovernmental Panel On Climate Change has stated that Scientific evidence for warming on the climate system is unequivocal and that there is a 95% probability to be the result of human activity [1]. Some of evidence that confirms this crisis is the increase of 2.12 degrees Fahrenheit of the planet's average surface temperature, the warming of the ocean of more than 0.6 degrees Fahrenheit, the average lost of 279, and 148 billion tons of ice per year on Greenland and the Antarctica, respectively. Other evidence include the rise of global sea levels of about 8 inches during the last century, and many other [1].

One would think that it is evident the urgent need to invest on green and climate friendly companies and industries, but we live in a country and a world where the influence and interests of some corporations on our government and policy making drastically affect our ability to invest on sectors of the industry which could positively impact our chances to successfully fight against climate change. Now more than ever it is crucial for the global economy to transition to climate friendly industries, and it is paramount for the biggest economies such as the US to play a key role in this transition.

3.2 Stock Market Understanding

Our attempt to investigate the stock market behaviour naturally requires a basic understanding of the stock market. More specifically, we focused on aspects which could lead us to narrow down certain public companies to track through their tickers. The value of a public company's stock can be retrieved by looking up their respective tickers. We discovered Exchange Traded Fund (ETF), which is a type of security that tracks an index, commodity, sector, or other asset, but can be traded like a stock. ETF can track a ticker as well. Market indicies proved very useful - they are hypothetical portfolios used to track specific segments of the market. They would include important tickers to keep track for that specific

market sector. Through ETF and clean energy indicies such as the S&P energy index, we were able to identify a number of companies we could focus on for our purposes. The companies we decided to include for our study are First Solar, Siemens Gamesa, Sunrun, Plug Power, Sunpower, and Plug Power. Another factor we used to choose these companies were their popularity and activity on Twitter, since this was essential for our work.

4 DATA UNDERSTANDING

Data Understanding phase of CRISP- DM Framework focuses on collecting the data, describing and exploring the data. This stage comprises of four key steps to understand the available data, and identify new relevant data in order to solve the business problem.

4.1 Collection of Initial Data

For this project, we collected three sets of data.

4.1.1 Twitter Data. The Twitter Data was obtained from the official Twitter API. We were able to obtain an Academic License which allows us to collect up to 10 million tweets per month. For our purposes, we collected around 1 million tweets for the five companies identified in the business understanding phase from 2011 to 2020 using the full archive end point. With the streaming API, we also collected around 1500 tweets for each company daily.

4.1.2 News Article Data. The News article data was collected using Bing News Search API. We used the Basic License, which gave 1000 requests per month. Each request gave up to 100 news articles, which could include duplicates from previous queries. The API also only retains articles up to one month old. This was naturally very limited. We were able to collect around 200 unique news articles per company for one month. Due to the limited nature of the news data we were able to collect, it proved difficult to use it in analysis.

4.1.3 Stock Prices Data. To collect the stock price data, we used the open-source Alpha Vantage API. It uses a ticker based search, allowing us to look up each company's stock directly. We collected data for each company with the longest duration allowed.

4.2 Description of Datasets

4.2.1 Twitter Data. The query returned by the Twitter API is a JSON file. Each object in the file has the following fields:

(1) Created at
(2) Tweet ID
(3) Author ID
(4) Tweet text
(5) Hashtags

4.2.2 News Article Data. The query returned by the Bing News Search API is a JSON file. It included a description of the search, and the list of articles returned by the query. Every article returned have the following fields:

(1) Item type
(2) Date published
(3) Title
(4) Description
(5) Provider
(6) URL

Not all data fields are always filled in, but the ones listed seem to be consistently available.

4.2.3 Stock Prices Data. The stock prices data for a certain company (ticker) given by AlphaVantage API is in the following format:

(1) Time Stamp
(2) Close value
(3) Open value
(4) High
(5) Low
(6) Volume

The time stamps values are in day, and the other fields are the corresponding values of the company for that day.

4.3 Exploratory Data Analysis

This phase of the project is a critical part of data understanding, it is exploring the data through plotting charts of various kinds with various combinations. Following types of insights can be achieved through plots/graphs. a) Spotting outlier values b) Observing trends of variables (increasing/decreasing) etc.c) Observing correlation and use fullness of variables in context and scope of the problem at hand

4.3.1 Stock prices - EDA. **Attributes' Predicting Value Assessment** As discussed above, from the API we used to collect the stock price provided various attributes regarding each stock which included:

(1) Closing price
(2) Opening Price
(3) Daily Highest
(4) Daily Lowest
(5) Volume
(6) Delta - derived attribute(|daily high - daily low|)

Out of all these attributes rather than using every single we one separately and then concluding if it is useful or not we on the other hand decided to apply Auto-correlation based analysis for time series which allowed us to approximate the predicting power of each of the given series based on the randomness it comprises of and eliminated the non-useful ones.

Autocorrelation, also known as serial correlation, it is the correlation of a series with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them and based on the results of this type of computed correlation we are able to approximate the randomness that a particular series comprises and discard it if there it is an underlying characteristic of it made obvious using the results.

To interpret the result, we should know that if the auto-correlation plot lines lie very close to zero then it can be said that the given series corresponding to one any of the attribute under consideration is random and hence not useful for the time series analysis. Performing this whole process on all of the given attributes we learned that the delta attribute is pretty useless as seen in figure 1 and out of the rest, all exhibit the same pattern. Therefore, using any one of it will suffice. We chose to use the "closing price" attribute series.

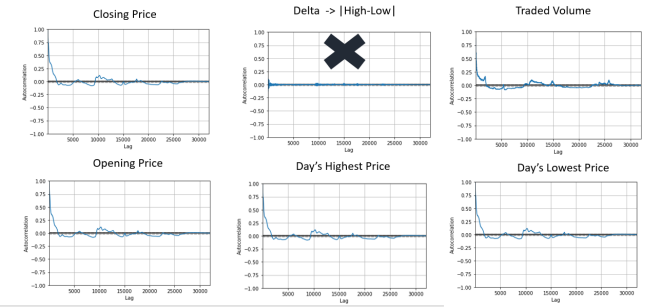

Figure 1: Stock Price Attribute Assessment - Predicting Power/Usefulness for time series Analysis

Closing Price Trend Plots Since we have decided to move forward with the closing price attribute only, we then explored the individual trend lines for the companies under consideration. This plot has two parts where Figure 2 shows the trend of every stocks closing price for the maximum data available and then figure 3 shows the same but a deep dive and closer look at the trend for a single latest year only. The granularity of figure 2 is yearly quarters where as for the second plot in figure 3 the granularity is of yearly weeks.

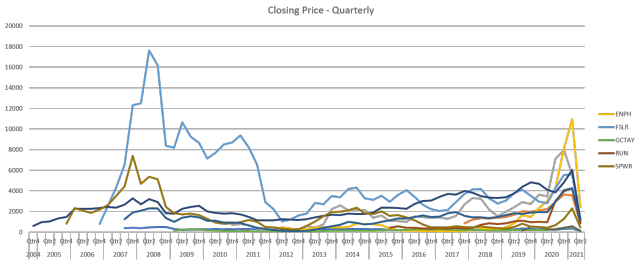

Figure 2: Stock Prices- Maximum Time Period for every company available

Figure 3: Stock Prices- 2019 - 2020

4.3.2 Twitter Data - Word Clouds. In the figures 4 and 5a we can see that the word cloud plots for the raw but cleaned twitter data. By raw we mean to specify that it not characterized by sentiment yet rather only split and segregated over individual companies. It was interesting to see what concepts overpowered each of the companies and did actually provide a very good idea of the most important strings attached to each particular company. A good example of it was for Siemens where we clearly saw visible focus on solar and wind energy whereas in the remaining one can easily conclude that there wasn't much of he talk around or relevant to wind energy. Another good use of this word-cloud analysis was that is gave us sound basis and reasoning to exclude Tesla's solar power company i.e "Solar City" because one could easily see it being overpowered by Elon Musk related and Tesla related stuff leaving the company itself to be quite insignificant.

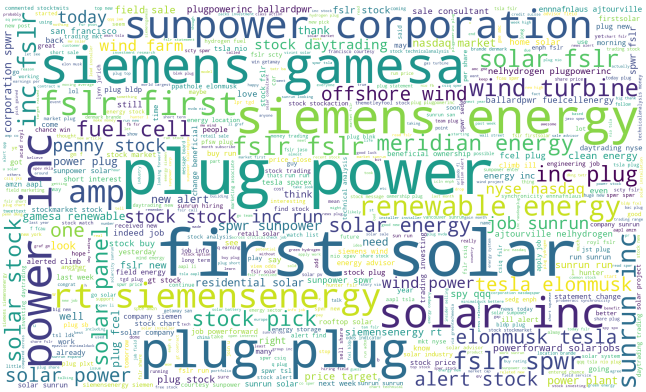

Figure 4: Word cloud - All Companies Combined

(a) First Solar **(b) Plug Power** **(c) Sun Power**

(d) Siemens **(e) Sunrun**

Figure 5: Twitter Data - Word Clouds

4.3.3 Author id - distribution. In this small analysis we applied methods for elbow analysis which sets the foundation and is usually used to identify the cut-off for the use of 80-20 rule as shown in figure 6. The 80-20 rule itself is characterized generally as 80% of contribution from 20% population or any data measures for that matter which in our case was for the idea that we were able to find the top 2600 tweet authors that were contributing in forming the 75% of twitter data that we collected. This analysis can be useful later when we make modifications to approaches we test for data collected where we can employ these results and collect data only for those 2600 authors or since we know they are sort of the influencers or top contributors that are forming the opinion of the crowd we start following their views directly on other domains and aspects as well. This will turn out to be useful later but first the scope of this project we have not made use of its discussed applications for now.

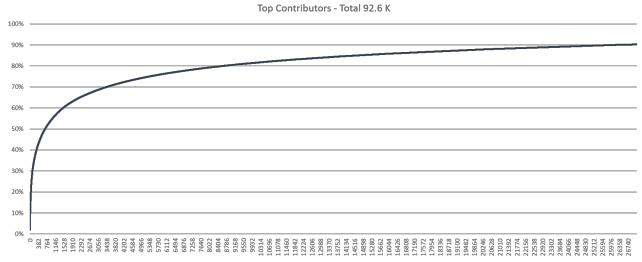

Figure 6: Tweet Author Importance - 80/20 rule cut-off identification

5 DATA PREPARATION

After having understood the data that we have collected and that which is available to us along with the sense of how we can make it useful for our particular problem we now move on to verify data quality and to remove errors from it so that we are then able to construct the intermediate forms and representations of it which are then finally integrated together to form inputs for the modeling step.

5.1 Data selection

5.1.1 Twitter Data. For Twitter data, we retained the following fields:

(1) Created at
(2) Tweet ID
(3) Author ID
(4) Tweet text

5.1.2 News Article Data. For news article data, we retained the following fields:

(1) Date published
(2) Title
(3) Description

5.1.3 Stock Price Data. For stock price data, we simply retained all data from the API as they are.

5.2 Clean Data

5.2.1 Twitter and News Article Data. To perform downstream tasks such as sentiment analysis, the text data from Twitter and news articles must be cleaned. For Twitter, the tweet text was cleaned, and for news articles, the title, combined into one field with the description, was cleaned. We cleaned the data with the following steps:

(1) Remove punctuation - Remove all punctuation from the sentences.
(2) Tokenization - Transform the sentences into a list of words that can be found in the sentence.
(3) Remove stop words - Remove all English stop words, which do not contribute to the sentiment.
(4) Stemming - Turn all words into their root (stem) form.
(5) Lemmatization - Turn all stem words into dictionary form.

We used Python to perform all these steps using NLTK as the primary library. We relied on the tutorial [4] to understand how to use this library and implement the required analysis.

5.3 Construct Data / Integrate Data

5.3.1 Twitter and News Article Data. Using the cleaned text data from above, for each entry in the Twitter and news article data we also created a field for sentiment. Using the NLTK Sentiment Intensity Analyzer, we found the sentiment for each body of text. These were used to construct time series in modeling.

5.4 Sentiment Analysis

The branch of data mining corresponding to use of machine learning algorithms to categorize various input samples of relevant text into the dominant and underlying contextual sentiment or notion that characterizes the tone of the text. A sentence "The movie was not bad at all" contains words like bad but the context and underlying notion of it contextually is rather a positive comment regarding it the subject under discussion, hence sentiment analysis makes use of ML techniques, applied strategically and efficiently to classify the text into certain categories of sentiment/nature of the text. Most commonly positive or negative sentiment. For our project we aimed to experiment and maximize the best approximation by going through various options available in the open-source community but ended up testing two sentiment analyzer tool options available and known as the best ones yet. We chose to employ some of the well established and well known pre-trained sentiment analyzer model available as open-source tools. These included

(1) Natural Language Tool Kit Sentiment Analyzer for python
(2) Stanford CoreNLP sentiment analyzer

NLTK. A powerful built-in machine learning operation to obtain insights from linguistic data. It is an open-source library available for python, it is designed to work with human language data making sentiment analysis a necessary tool of it. It provides the complete set of tools required for contextual mining and hence very popular and known for its performance among the text mining community and industry. To be specific we used the **SentimentIntensityAnalyzer** model instance followed by the polarity score measures it computes.

Returned Polarity measures per text sample:

(1) Positive Sentiment - between 0 and 1 with zero being lowest positivity score and 1 being highest
(2) Negative Sentiment : between 0 and 1 with zero being lowest positivity score and 1 being highest
(3) Compound Sentiment: Less than zero signifies negative sentiment and greater than zero signifies positive sentiment. 0 score value here corresponds to neutral sentiment
(4) Neutral Sentiment

We chose to construct time series for Positive, Negative and Compound sentiments to maximize the potential for finding some useful correlation with any of the three. Neutral Sentiment measure was redundant hence not used further.

Stanford CoreNLP. CoreNLP enables users to derive linguistic annotations for text. Although it is a very powerful tool, but following are a few issues that came-forth whiles using it which made us choose and thus opt for the NLTK sentiment analysis and let go of this tool.

(1) Slow speed
(2) Extensive error debugging required to pin point the issues causing failure of a long run and then restarting the analysis, we did try to do it but it ended being an extensive and extreme wastage of time and resource power of our machines.

5.4.1 Stock Price Data. Using the "High" and "Low" field in every record, we created a new field "Delta"

$$Delta = |High - Low|$$

"Delta" returns the absolute value of the change in stock price for a specific stock for that specific day. It can be thought of as a reduction function for the daily high and daily low.

6 MODELING AND EVALUATION

An integrated model comprising of multiple modeling sub - components has been used to formulate, investigate and analyze the underlying relationship between the stock prices for the top five green industry companies from our list against a few "wisdom of crowd" measures constructed using twitter data; the most significant and useful ones ending up to be the twitter sentiment and entity popularity using tweets. After individually completing the sub-components; the lag analysis and Granger Causality analysis is performed to get a sense of model behavior and its potential to facilitate in making data driven decisions for investment in this industry and particular companies. We make use of the lag analysis with some tweaking and variations, in this section we present in detail some aspects of the approach and techniques we have employed in order to achieve the identified goal.

6.1 Explore Data Analysis 2 - Explore sentiment data individually

To get an idea of the big picture from the sentiment analysis results, we plotted a summary of the positive, negative, and compound Twitter sentiment from 2011 to 2021 in figures 7 to 9, respectively. These plots are specially useful to distinguish an evident anomaly as shown.

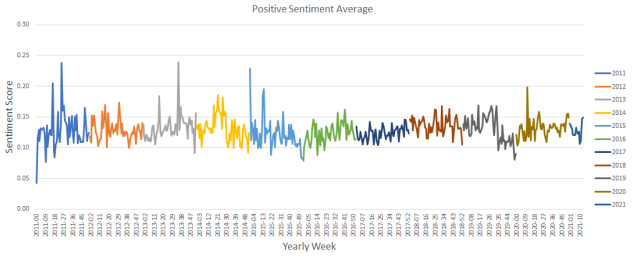

Figure 7: Positive Twitter Sentiment - Year Wise

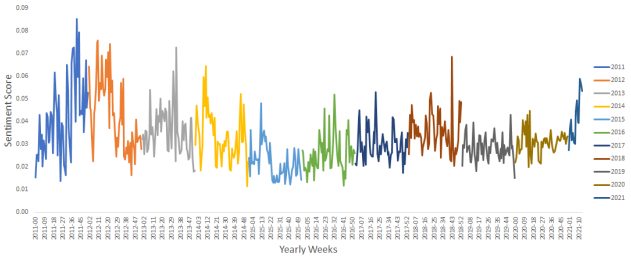

Figure 8: Negative Twitter Sentiment - Year Wise

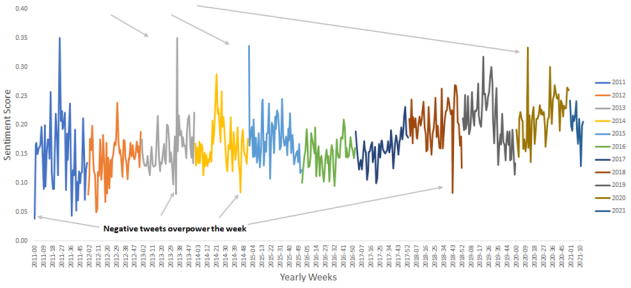

Figure 9: Compound Twitter Sentiment - Year Wise

6.2 Twitter Popularity and Stock Prices Series

After the initial data exploration. We plotted the Twitter popularity and the stock prices time series. We define the Twitter popularity simply as the daily count of tweets for the respective company. The plots in figures 10 and 14 show these time series. To the reduce the noise of the daily tweets (popularity), we smooth the series taking the mean of a rolling window of 10 days.

We can observe, as depicted with the red boxes that there is certainly some similarity between the Twitter popularity and the closing stock prices with certain lag. This observation will be investigate using Granger causality, to determine whether the count of daily tweets Granger causes the stock prices.

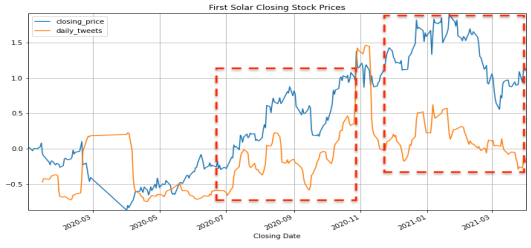

Figure 10: First Solar Closing and Popularity Series

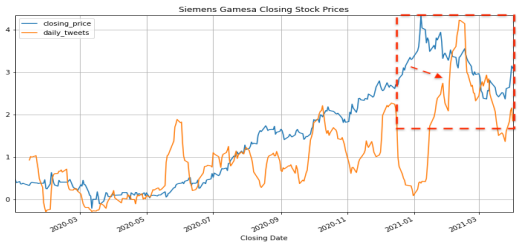

Figure 11: Siemens Closing and Popularity Series

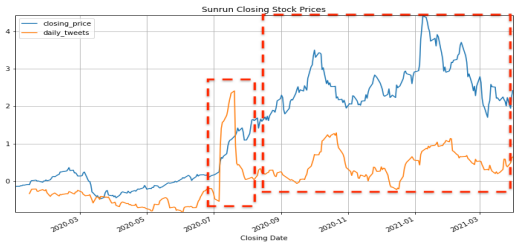

Figure 12: Sunrun Closing and Popularity Series

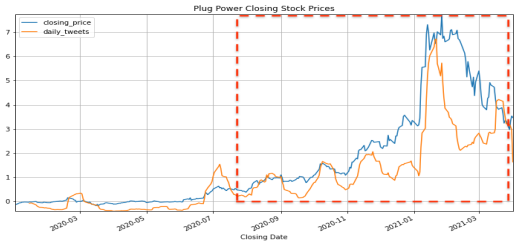

Figure 13: Plug Power Closing and Popularity Series

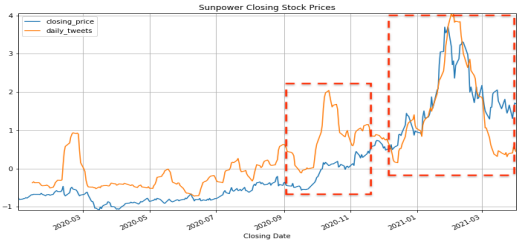

Figure 14: Sunpower Closing and Popularity Series

6.3 Compound Sentiment and Stock Prices Series

We followed a similar approach to that of the previous section the compound sentiment and closing stock prices. Nevertheless, just by inspection, there is not many clear similarities in the two series. This observation will be investigated as well in the next sections.

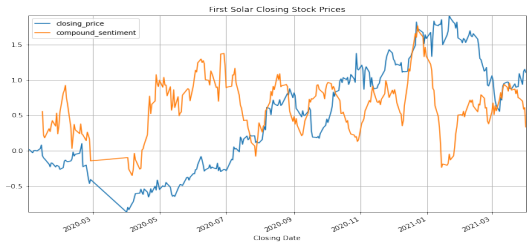

Figure 15: First Solar Closing and Sentiment Series

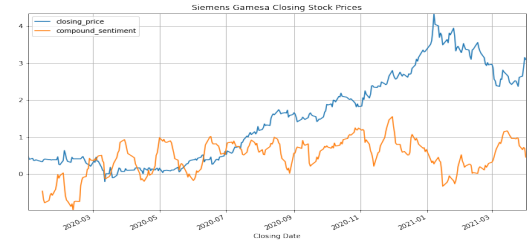

Figure 16: Siemens Closing and Sentiment Series

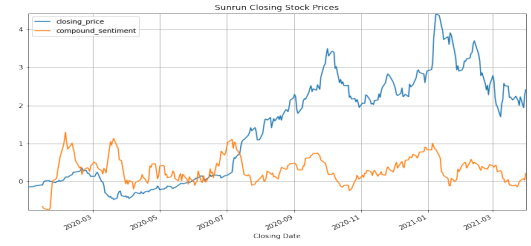

Figure 17: Sunrun Closing and Sentiment Series

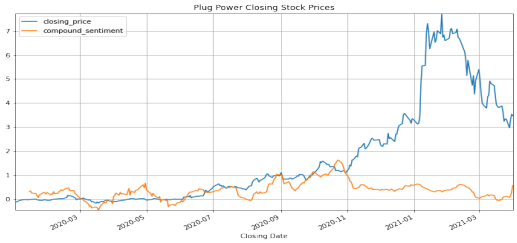

Figure 18: Plug Power Closing and Sentiment Series

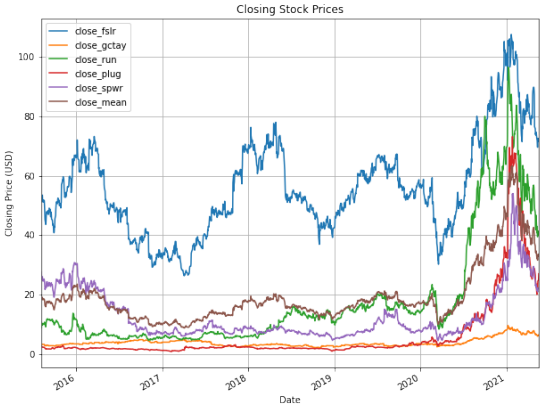

Figure 20: All Stock Prices - 2015 to 2021

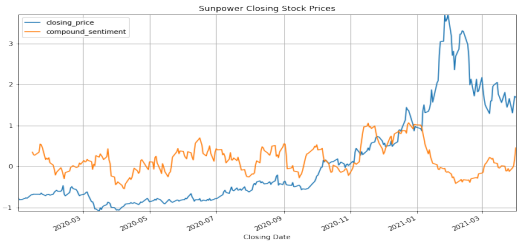

Figure 19: Sunpower Closing and Sentiment Series

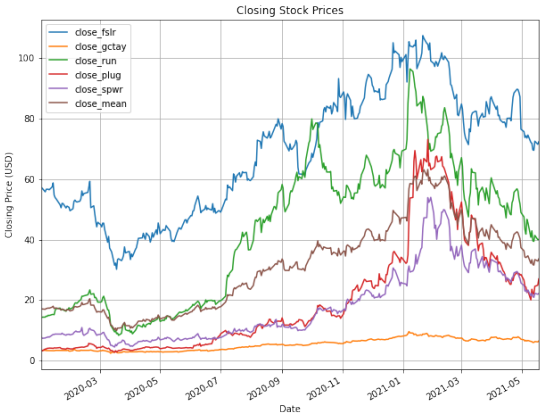

Figure 21: Significant Pace of Volatility in all green energy companies in later half of 2020

6.4 Lag and Correlation Analysis

A distributed lag model is a model for time series data in which a regression equation is used to predict current values of a dependent variable based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable.

"After extensive lag analysis and simplifying the convoluted and counter intuitive results that we were getting, we were able to conclude that the green energy industry as a whole has just picked up the pace. This has been observed in the data, as illustrated in figures 20 to 21, and is transforming the usefulness of our approach from long term uniformity that we witnessed of stock prices for these companies in the open market into a much shorter term predictive capability; very slightly now (given the recent awareness and shift, July 2020 on-wards) but very surely so. Given a few more years for this industry to mature, this same approach we have presented in the paper will be capable of short term predictive usefulness for driving short term insightful decisions as well. "

6.4.1 Popularity Vs. Closing Price. In this section, we present the various combinations we tested for the lag analysis of each company against the positive, negative and compound sentiments (figures 22 to 33) respectively and discuss what insights it provides us for figuring out the underlying correlation. For every company, we plotted a lag analysis initially for a max lag of 100 against stock-prices and each sentiment(Positive, Negative and Compound) but realized that the distribution of correlation is somewhat constant over this time and it only slightly varies towards the positive lag values.

It was difficult to comprehend why this was but then plotting lag analysis with max-lag values of greater time periods like 500 days

and 1000 days we saw a better distribution that made sense. This allowed us to understand that the current relationship between stock prices and sentiments is stable and slow-changing. The market gap is easily 100 days which sounds pretty off but when analyzed against the raw stock prices we had for the companies and their insignificance and awareness in the market people and investors validated the fact that the very last months of 2020 i.e to be specific from July 2020 on-wards we see a significance movement in the majority companies' stock prices which reinforces and validates the consistent distribution of correlation in lag analysis.

That consistent part highlights the idea that people started talking about these green companies earlier and have been for a while but it never translated into action of investment that might be significant in contribution and growth of the stock itself, the true investment action although volatile in nature such that it is following cycles of rise and fall in the time period after July 2020 to latest days although does prove to now give us some useful material for varying correlation.

The uniformity of the stock prices in the majority of the time overpowered and not much useful and significant relationship in the short term cut-offs rather long term correlation is pretty much meaningful and so what we concluded overall from all the analysis that will now be listed and below we concluded that the green-investing energy is not quite volatile at the moment and the decision making can be done for the longer term rather short term given the current talk in the town and awareness in the audience we have taken under consideration.

Simply put we have identified that there has been some shift in the attitude of the investors towards this green industry from July 2020 on-wards although volatile in this shorter time period but still now people are talking about it. Whereas for the majority preceding chunk of time there was uniformity in the green industry not making analysis like ours useful but this shift that we have been able to recognize in the attitude of people towards this industry and its reinforcements from the data validates that giving some more time to this industry i.e a few years or so it will be very useful to have an analysis replicated as ours which would then start indication short-term correlation and decision making.

So finally setting it straight the current analysis leans towards helping us gain long-term insights on the stocks and if we give some time to this green industry we can reuse this same approach and we will most likely be able to present short-term insights gained with a reduced market gap. Right now our market gap is in unit of years basically rather than days but once the industry grows further given the trend we started observing from later half of 2020 it should easily reduce down to weeks and days etc. Any day we choose as optimal lag value for now between 0 - 100 or so will be giving us the same insights so the daily volatility becomes insignificant and hence we can get a longer term picture for now because what correlation we might get for a day 200 days away from the chosen one is where we start seeing decline rather than the very next day. Our lag analysis basically done on yearly or half year granularity would have given us something we expect as the change with smaller number of point until max lag. So imagine we have the lag analysis chart below and every line pair of consecutive lines/points are time wise at a distance of 6 months.

6.4.2 Discussion for Lag and Correlation analysis . All except First Solar in figure 20 recently gained the pace and that was an interesting find which made sense to the uniformity of the lag analysis over shorter chunks of time. Where a lag of one or hundred would end up giving the same results as compared to obvious delta and declining correlation in increasing lags significantly to greater lag ranges. Figure 20 illustrates that something is going on the crowds very recently that is reflected in people now buying and selling more of the green energy company stocks as opposed to the non-popularity and uniformity they were facing in the previous years. This phenomenon is visible in figure 20 and 21 very clearly. The lag analysis' slight variations on one half of the plot in the 100 lag category reinforces this as well. The position of this half decreasing curve formed in the following lag analysis will vary depending on whether we are plotting the price against positive or negative sentiment. As ample for two of the companies is attached below with the rest available in appendix.

Lag Analysis For First Solar

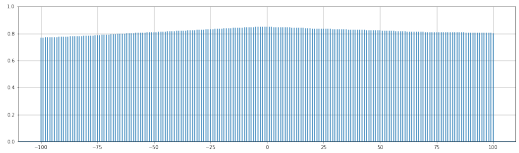

Figure 22: First Solar - Positive Vs Closing Price - Max Lag =100

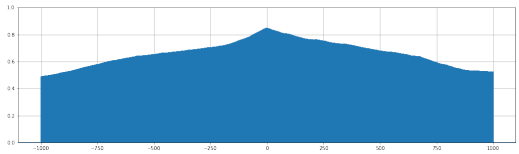

Figure 23: First Solar - Positive Vs Closing Price - Max Lag =1000

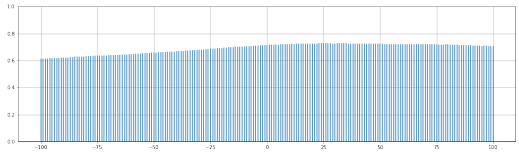

Figure 24: First Solar - Negative Vs CLosing Price - Max Lag =100

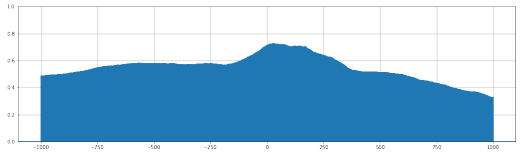

Figure 25: First Solar - Negative Vs Closing Price - Max Lag =1000

,

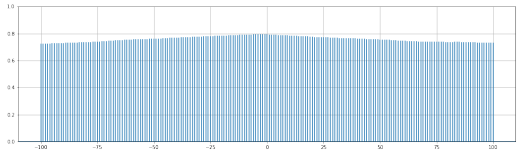

Figure 26: First Solar - Compound Vs CLosing Price - Max Lag =100

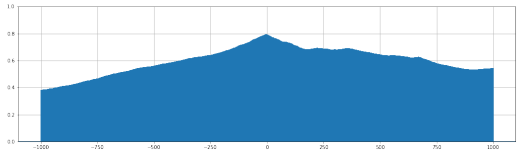

Figure 27: First Solar - Compound Vs Closing Price - Max Lag =1000

Lag Analysis For Sun Run

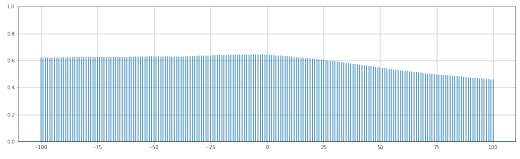

Figure 28: Sun Run - Positive Vs Closing Price - Max Lag =100

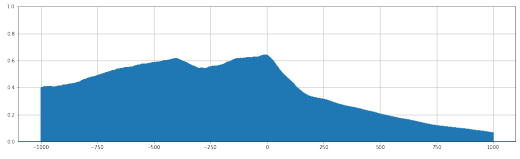

Figure 29: Sun Run - Positive Vs Closing Price - Max Lag =1000

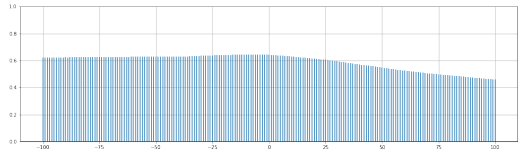

Figure 30: Sun Run - Negative Vs CLosing Price - Max Lag =100

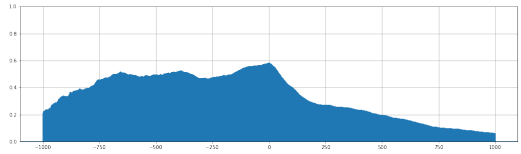

Figure 31: Sun Run - Negative Vs Closing Price - Max Lag =1000

,

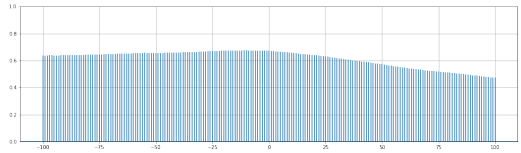

Figure 32: Sun Run - Compound Vs CLosing Price - Max Lag =100

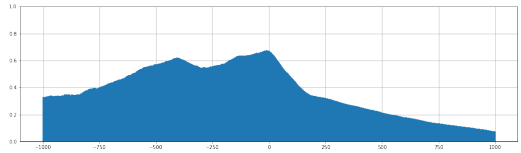

Figure 33: Sun Run - Compound Vs Closing Price - Max Lag =1000

6.5 Granger Causality

The Granger causality results are presented in figures 34 to 38. For each of the companies we did four F-tests. We used compound, positive, negative sentiment, as well as daily tweets to determine whether each series Granger causes the closing prices. The determine whether one of these series Granger causes closing prices, we looked for lag values which have a very low p value ($p < 0.05$). Those lag with their respective p and F values are shown in the figures 34 to 38.

For each company, different series Granger causes the closing prices at different lags so we cannot make a definitive claim about the effectiveness of one series with respect to another one for all of the companies. Nevertheless, we can observe that the amount of daily tweets is the most successful at Granger causing the closing prices except for First Solar. It is worth noting that the figure only some of the lags with $p < 0.05$.

We can indeed check that the lag values output by the F-test do make sense with the time series plots in figures 10 to 19. Specifically, we can observe that the lag values do match with the shift of the daily tweets and the closing prices, indicating that twitter popularity does predict up to a certain extent the closing prices. In general, we can also see that sentiment scores shows to be less effective at Granger causing values of closing prices, and that for the companies Sunrun, Plug Power, and Sunpower, the sentiment scores is not relevant at all for this scenario.

Company	FSLR - First Solar	
Maxlag	100	
Series 1	Closing price	
Series 2	Compound Sentiment	
F	p	Lag
5.6361	0.01766	1
3.2806	0.03776	2
Series 2	Pos Sentiment	
F	p	Lag
-	-	-
Series 2	Neg Sentiment	
F	p	Lag
8.0898	0.004487	1
4.6287	0.00984	2
3.5041	0.01481	3
Series 2	Daily Tweets	
F	p	Lag
-	-	-

Figure 34: Granger Results - First Solar

Company	GCTAY - Siemens	
Maxlag	100	
Series 1	Closing price	
Series 2	Compound Sentiment	
F	p	Lag
1.4857	0.0372	33
1.4736	0.03647	35
1.4565	0.03918	36
Series 2	Pos Sentiment	
F	p	Lag
-	-	-
Series 2	Neg Sentiment	
F	p	Lag
5.6852	0.01718	1
3.3047	0.03686	2
1.3578	0.049065	50
Series 2	Daily Tweets	
F	p	Lag
3.2179	0.006707	5
2.80645	0.01006	6

Figure 35: Granger Results - Siemens

Company	Sunrun - RUN	
Maxlag	100	
Series 1	Closing price	
Series 2	Compound Sentiment	
F	p	Lag
-	-	-
Series 2	Pos Sentiment	
F	p	Lag
-	-	-
Series 2	Neg Sentiment	
F	p	Lag
-	-	-
Series 2	Daily Tweets	
F	p	Lag
4.6766	0.00945	2
3.8605	0.009145	3
3.195828	0.01264	4

Figure 36: Granger Results - Sunrun

Company	Plug Power - PLUG	
Maxlag	100	
Series 1	Closing price	
Series 2	Compound Sentiment	
F	p	Lag
-	-	-
Series 2	Pos Sentiment	
F	p	Lag
-	-	-
Series 2	Neg Sentiment	
F	p	Lag
-	-	-
Series 2	Daily Tweets	
F	p	Lag
4.3354	0.004695	3
10.5208	0	5
7.87492	0	6

Figure 37: Granger Results - Plug Power

Company	Sunpower - SPWR	
Maxlag	100	
Series 1	Closing price	
Series 2	Compound Sentiment	
F	p	Lag
-	-	-
Series 2	Pos Sentiment	
F	p	Lag
-	-	-
Series 2	Neg Sentiment	
F	p	Lag
-	-	-
Series 2	Daily Tweets	
F	p	Lag
6.85488	0.008892	1
2.72826	0.0425	3
1.62902	0.03248	22

Figure 38: Granger Results - Sunpower

6.6 TF-IDF: Term frequency–inverse document frequency

We also performed *TF_IDF* to get key word extraction, the results are as below and this can be further used to develop more insightful analysis but for this project due to limited time we plan on making use of it creatively for extension of the project to new aspects.

Top 20 key words for each company have been computed but it varies depending on the importance returned, they might be and usually were less than 20.

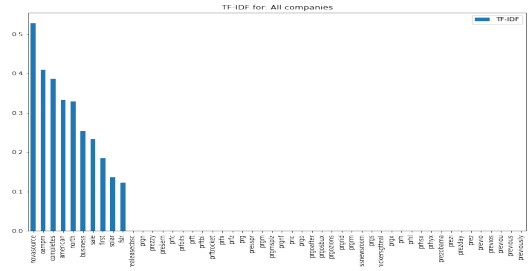

Figure 39: TF-IDF - All companies

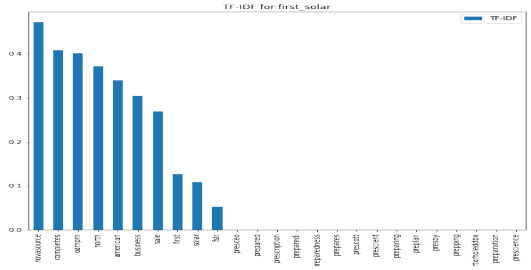

Figure 40: TF-IDF - First Solar

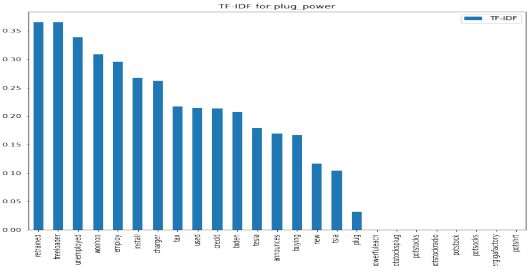

Figure 41: TF-IDF - Plug Power

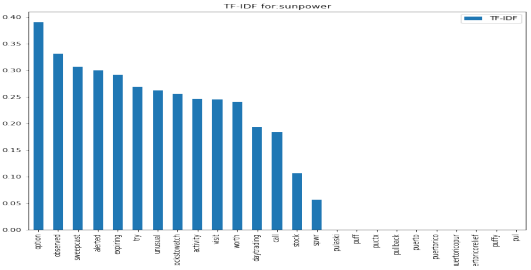

Figure 42: TF-IDF - Sun Power

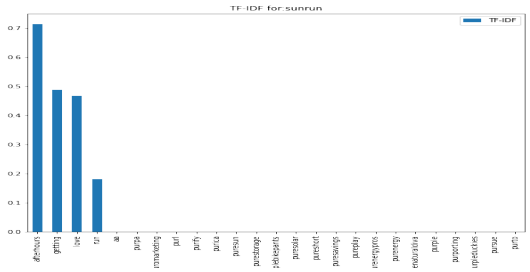

Figure 43: TF-IDF - Sun Run

7 CONCLUSION

After doing extensive work on implementing the methodology we had in mind for solving the identified problem statement of **"Using Twitter and news article data to predict stock market prices for green energy companies"** we have understood the underlying relationship to some extent after several iterations.

An interesting learning point was that due to pre 2020-July period for the green energy industry there was not much going on and so very little margin to explore the highest potential of data driven decisions but the post July-2020 times did put a lot in perspective and so we also see that towards the end of the time series there is significant similarity between the closing stock prices and the popularity and sentiment measures.

We conducted many Granger causality tests for all the companies to investigate the predictive properties of twitter popularity and sentiment analysis to predict the behaviour of the closing stock prices of the mentioned companies. After doing this tests, we were able to confirm that the lag values output by the F-test match the respective time series. Specifically, we can observe that the lag values do match with the shift of the daily tweets and the closing prices, indicating that twitter popularity does predict up to a certain extent the closing prices. In general, we can also see that sentiment scores shows to be less effective at Granger causing values of closing prices, and that for the companies Sunrun, Plug Power, and Sunpower, the sentiment scores are not relevant at all for this study.

For future studies, we would like to expand the list of companies in the green industry sector as well as other industries to investigate how well our conclusions generalize to other scenarios. We would also like investigate other alternative data sources such as satellite imagery, or other weather related data that could have a potential impact on the performance of renewable energy companies. We would also like to use other features of the Twitter API such as likes, or other engagement actions that could enrich our models.

REFERENCES

[1] 2021. Climate Change Evidence: How Do We Know? https://climate.nasa.gov/evidence/

[2] A. Bari, P. Peidaee, A. Khera, J. Zhu, and H. Chen. 2019. Predicting Financial Markets Using The Wisdom of Crowds. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*. 334–340. https://doi.org/10.1109/ICBDA.2019.8713246

[3] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR* abs/1010.3003 (2010). arXiv:1010.3003 http://arxiv.org/abs/1010.3003

[4] Rageeni Sah. 2018. Text Data Cleaning - tweets analysis. https://www.kaggle.com/ragnisah/text-data-cleaning-tweets-analysis