COLUMBIA UNIVERSITY

COMS E6998: Machine Learning and Climate

Final Project Report - Spring 2022

Search for a Connection between Energy Demand and Media Sentiments

Sunjana Ramana Chintala
Uni : sc4921

Abstract

New York's energy bill fluctuations are a growing concern of the people. Due to the unpredictable changes in energy demand, suppliers struggle to allocate sufficient resources for energy production. This imbalance causes energy bill rates to surge which leads to residents not being able to plan their household budgets accordingly. The goal of this project is to build a robust Time Series Forecasting Model that predicts Energy Demand effectively. In addition to this, the project also attempts to investigate the reason behind the anomalous highs and lows in energy demand by studying social media sentiments. Data for Energy Demand is collected from New York ISO website. ARIMA model is built for Time Series Forecasting. Twitter Data is streamed for anomalous data and topic Modeling techniques like LDA and GSDMM are used for the analysis of textual data.

Keywords: **NLP, Topic Modelling, Energy Demand, Time Series Forecasting, Machine Learning, Climate**

Contents

1 Introduction

New York's energy price hike is a pressing concern of the residents. According to the article by Sophie Mellor(1), 1.3 million New Yorkers fall behind on energy bill payments. During the winter months, the utility bills for the residents skyrocketed by 50%, which has led to an increase in the number of overdue energy bills for the people. Raising concerns of people having service termination for gas and heat has pivoted to an all-time high.

When the suppliers were enquired about the reason behind these exorbitant prices, they blamed it on the wholesale power prices which are driven by generators. The energy bills are decided based on two factors namely delivery and supply charges. While delivery charges are always constant, supply charges depend upon market supply and demand. Effective demand forecasting is crucial to help suppliers decide upon how much electricity would be made available on a particular day. Energy Demand Forecasting also allows residents to foresee any extreme utility bill increases. This would allow people to plan their household budgets accordingly.

With that being said, the article also pointed out that the rise in energy prices during the winter months has been driven by higher winter demand. Though this point of view is widely accepted, there is little evidence to support the assumption. There are several days in a year when the energy demand has anomalous highs or lows. The reason behind these anomalies is not known. To understand what other reasons there could be, the power of social media can be utilized. Social Media is one such avenue where people tend to voice all their opinions. By analyzing the most trending topics on days of very high or very low demand, the actual reason behind these spikes can be derived. A more concrete explanation of the jumping prices can be understood. The goal of this project is to effectively forecast energy demand prices and also provide reasoning for the fluctuating prices through social media sentiments.

2 Related Work

2.1 Time Series Forecasting

Data Mining helps in investigating patterns present with large datasets to predict future outcomes. There is a broad range of techniques within data mining, that have a variety of advantages when they are applied to certain problems(2). Time Series Forecasting as one of these techniques helps in leveraging numerous data points in time to predict data-driven outcomes which allow decision-makers to derive value from historical data. One of the most interesting and widely used applications of time series forecasting is in the financial markets to predict stock prices. ARIMA(Auto Regressive Integrated Moving Averages) is the most seasoned approach used in this regard. Plenty of research has been done on how ARIMA alone and with the combination of other Machine Learning models like XGBOOST, Neural Networks, LSTM etc has aided decision-makers to drive the stock price index higher(3) (4) (5). This concept has also been applied to the Energy Sector to Forecast Energy Demand. Effective Load forecasting is an important topic of interest while studying power systems, especially when considering the changing climatic pattern and increased renewable energy penetration of the grid. Therefore, in this project, ARIMA is used for Energy Demand Forecasting on New York City's Load Data(6).

2.2 Twitter Streaming/ Mining

The concept of data streaming has evolved in the big data era. Streaming data is collected on a variety of topics and the public sentiments are derived. Ever since Twitter has made its data public, Twitter streaming has been used in a variety of scenarios to conduct research namely receiving for threat monitoring, analyzing trends in a certain region etc (8). By harnessing this power of Twitter streaming, public sentiments on days of high ad low demand can be investigated(7).

2.3 Topic Modeling

Topic modelling is a popular NLP technique that is used to analyze twitter sentiments. Latent Dirichlet Allocation(LDA) is used to determine the distribution of trending topics on Social Media(9). The simplicity of LDA is what marks its prominence as the most effective modelling technique. Experimental results have validated its effectiveness on textual data. However, for tweets which have shorter length text, GSDMM was found to infer a greater number of clusters by maintaining a balance between completeness and homogeneity. It was also found to cope with the sparse and high-dimensional problem of short text. In this project, both LDA and GSDMM are used for analysis(10).

3 Overview

The following figure gives a high-level outline of the entire project.All code for the project as been written in Python. Google Cloud Platform(GCP) and Google Colab are used for Backend support.

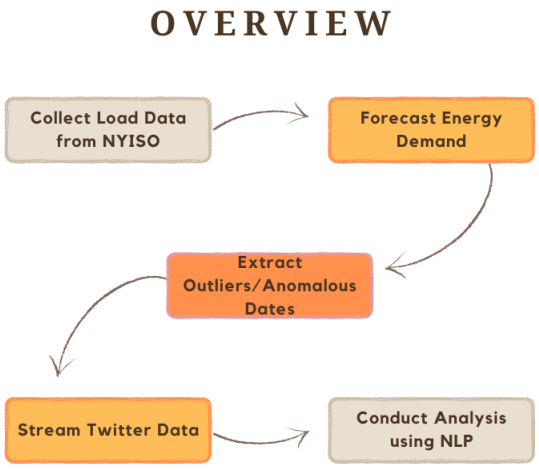

Figure 1: Outline of the Project

4 Data

In this project, three different kinds of datasets are used. They are as follows:
1.) Energy Demand Forecasting Data

Energy Load Data is collected from the EIA website through their open data API. Energy Information Administration (EIA) is a part of the U.S. Department of Energy. The U.S. Energy Information Administration (EIA) is the primary agency of the U.S. Federal Statistical System. They are responsible for collecting, analyzing, and disseminating energy information and are involved in policymaking. EIA website has APIs for a variety of data on coal, petroleum, natural gas, electric, renewable and nuclear energy. This dataset is in a time series format with data points from 2015 to 2022.

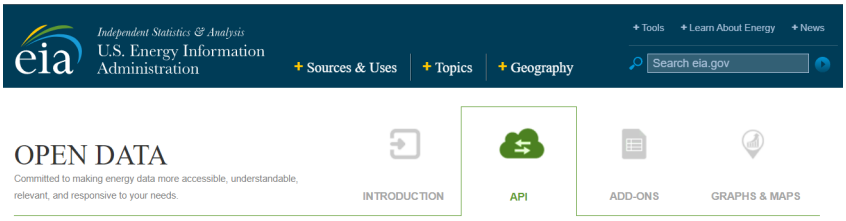

Figure 2: EIA Website

2.) Outlier Data

Outler Data is collected from the original time series dataset by establishing a threshold which gives a collection of dates. More information about the process will be discussed in the Methodology section.

3.) Streaming Data

To stream Twitter Data, the best tool for use is the Twitter API. However, this API does not allow streaming Twitter data that is more than a month old. The EIA load data consists of dates from 2015 onwards. Therefore to stream tweets from 2015 onwards, an alternative is needed. Other libraries like Tweepy, GetOldTweets2, TWINT etc are useful but do not fully serve the purpose as they either consume a lot of time or are expensive. Therefore, a viable tool was required.

"SNSCRAPE", a powerful scraping tool, helps in this regard. This tool can be used for obtaining twitter data on users, user profiles, hashtags, searches, threads, and list posts. Using this tool, tweets with high engagement are streamed to places in and around New York.

```
get_twitter_data(top_start_dates, top_end_dates, path )
Say it with me kids: calling a racist “racist” is NOWHERE NEAR the same level of bad as being racist yourself @nytimes
@emoturkeynugget it’s near new york n pennsylvania n other states tht i dont rly care about snjsjsjs
@AUkeibro Donald Trump is the son of an immigrant mother and an immigrant Grandfather. Mary Anne MacLeod (1912–2000) was
Summertime in New York is experiencing bugs falling from the heavens (or depths of hell?) and landing on or near you at
I'm old enough to VIVIDLY remember NYC in the 1970s and I'm still surprised by what I see every day near my office on W
A New York real estate developer and a veteran biotech investment firm are teaming up to build a lab on First Street, n

https://t.co/lMV0h1wnU9
A New York Central FA-FB-FB-FA set in lightning stripes lead a westbound freight near Bergen, New York in June of 1960.
New York's favorite chunky, gooey cookies (you know, the ones that break the internet every few years) could be coming
@max_tweedie After visiting *very gay* cities, New York for World Pride, then coming back to DC where the rainbow flags
```

Figure 3: Twitter Data

5 Methodology

5.1 Energy Demand Forecasting

After collecting the load data, a series of steps are performed to structure data in the desired format.

Step 1: Data Pre-processing The load data consists of two variables: Date and Demand(MWh). To ensure that the data collected is usable for all types of analysis, the Monthly and Yearly columns are extracted by calculating the Mean value of Energy Demand. These are stored as separate data files. Now, the original dataset is split into three datasets: Daily Data, Monthly Data and Yearly Data

Step 2: Data Cleaning Next, the datasets are cleaned. Any missing values or null values are removed to ensure smooth processing.

Step 3: Train-Val Split

In a regular scenario, a certain split ratio is declared and data is split accordingly. However, since the temporal aspect of the data needs to be preserved, the data is split manually without any such declaration. Data from the Year 2015 - The Year 2021 is considered to be training data and the remaining is stored as Validation Data.

Step 4: Data Visualization

A glimpse of the time series data can be seen below.It appears that the data is stationary and there is no trend present.

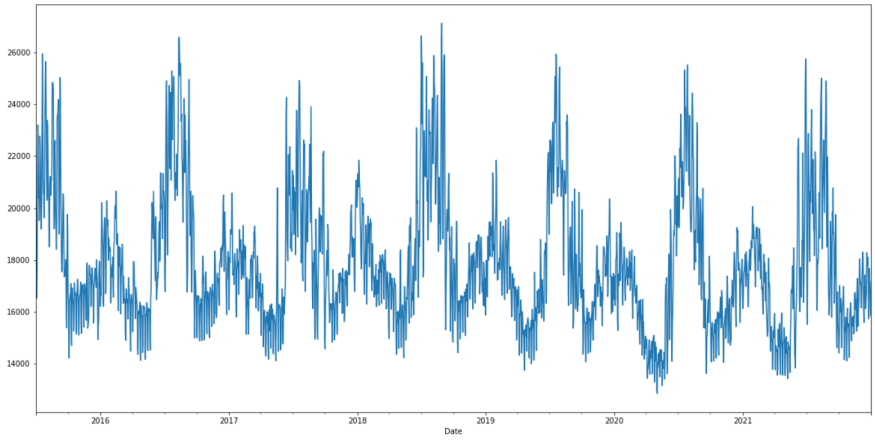

Figure 4: Demand v/s Time

The histogram representation of Energy Demand Data is plotted below:

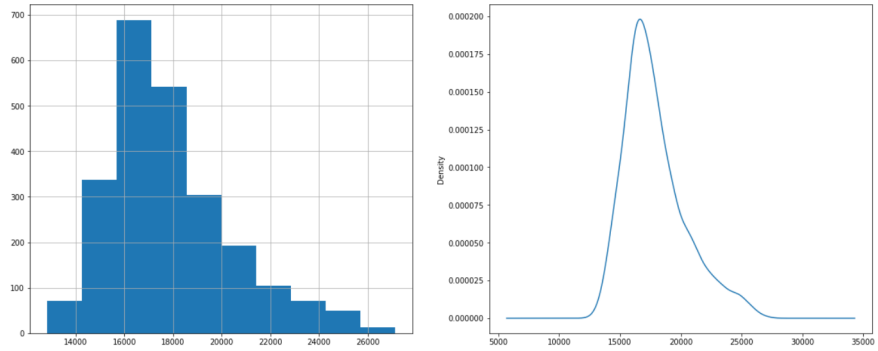

Figure 5: Distribution of Demand

Step 5: Fitting the Data into a Baseline Model

A baseline model allows us to get a rough estimate of how well the actual model is performing, on whether the performance is good or bad. It ensures a mode of comparison for the final model. In this case, the persistence algorithm is used to achieve baseline scores.

To implement the persistence algorithm, the univariate dataset is transformed into a supervised learning problem. Two columns are created by considering the difference in time. Training and test data is initialized. The predicted values are stored in a list.

RMSE is used as an evaluation metric. The RMSE value of the baseline model is obtained to be 1072.8 approximately. The final model that is constructed should have an RMSE score lower than the baseline RMSE score.

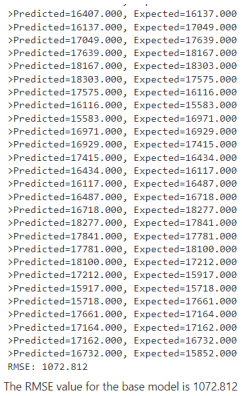

Figure 6: Baseline Accuracy Score

5.1.1 Time Series Forecasting using ARIMA

To perform Time Series Forecasting with ARIMA, the following conditions need to be satisfied:
1) The Mean should be constant throughout the data
2) Variance should be constant
3) There should not be any Seasonality or Trend within the data.

To check if the data is stationary or not, ADF(Augmented Dickey-Fuller Test) is used. The following are the results of the ADF test:

```
ADF Statistic: -4.792430
p-value: 0.000056
Critical Values:
        1%: -3.433
        5%: -2.863
        10%: -2.567
```

(a) ADF 0

```
ADF Statistic: -9.247085
p-value: 0.000000
Critical Values:
        1%: -3.433
        5%: -2.863
        10%: -2.567
```

(b) ADF 1

Figure 7: ADF Statistic for 0 and 1 differencing

Since ADF statistic(p-value) is less than 1% Critical Value for both 0 and 1 differencing, we can conclude that the data is not stationary. Therefore some differencing factor has to be applied to make the stationery.

Applying ARIMA:

ARIMA comprises three terms Auto-Regression, Integrated and Moving Average. To use the ARIMA model, the values of all these three components namely p, d and q would have to be declared.

1) Auto-Regression: Regression is nothing but predicting the future based on past values. Usually, Regression consists of two or more independent variables which determine the values of the Dependent Variable. Since Time-Series data only consists of one variable, Auto-Regression(AR) model is used here. Auto Regression means predicting future values in time based on the order of the previous values.

To determine the Auto-Regressive factor(p), it is important to see when the Autocorrelation plot has repetition in values.

The following are the Autocorrelation plots for Daily Data, Monthly Data and Yearly Data.

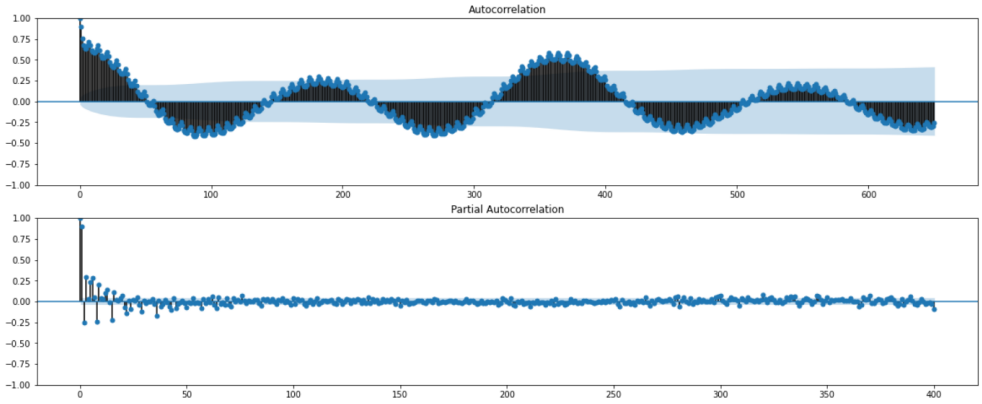

Figure 8: Daily Data Autocorrelation Plot

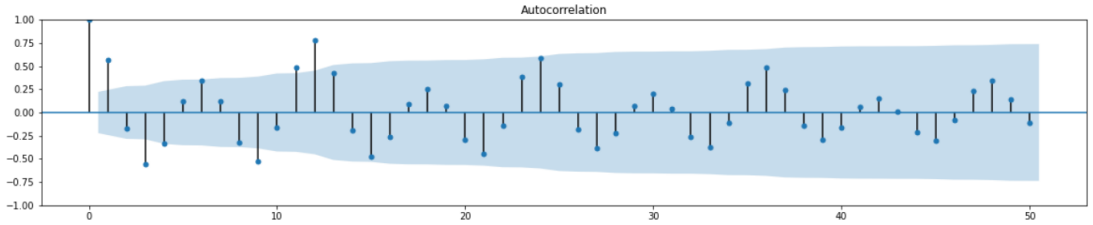

Figure 9: Monthly Data Autocorrelation Plot

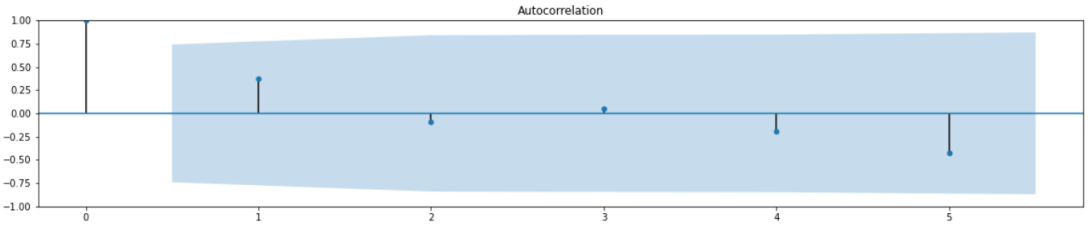

Figure 10: Yearly Data Autocorrelation Plot

For Daily Data: Here we can see some correlation when the p-value is around 365 Also, some inverse correlation is observed at 100 and 280 Let us now visualize using monthly data

For Monthly Data: At p = 3, some inverse correlation is observed At p = 11 and 12, some positive correlation is seen This shows that Today's demand is highly correlated with last year's demand on the same day.

For Yearly Data: This does not give much insight. Now that we know that the p-value is more significant in Monthly Data, we can use this data for further analysis. There are three possible p values through visual inspection namely p = 3, 11, and 12.

2) Integrated(d) :

We know from ADF tests that the data is not stationary. To achieve stationarity, we use the value of d. Using d=1 we would simply subtract the next value from the previous value. Differencing would eliminate any trend present in the data. Depending upon the nature of the data, first, second or third order differencing is used on the data. From visual inspection, it looks like first-order differencing would make the data stationary.

3) Moving Averages(q):

Moving Averages means using previous errors to make future predictions. Previous errors or lags are used to determine what the order of the model could be.

Since we now have a brief idea of what the p and d values could be, the value of q can be found through hyperparameter tuning:

From hyperparameter tuning, the best value of (p,d,q) is found to be (11, 1, 2). The RMSE value when these values are given as input to the ARIMA model is 916.859 which is better than that of the baseline score. The Model summer of the fitted values is seen below

```
ARIMA(6, 1, 0) RMSE=1367.733
ARIMA(6, 1, 1) RMSE=1373.337
ARIMA(6, 1, 2) RMSE=1233.712
ARIMA(7, 0, 0) RMSE=1329.826
ARIMA(7, 0, 1) RMSE=1340.036
ARIMA(7, 0, 2) RMSE=1195.220
ARIMA(7, 1, 0) RMSE=1372.742
ARIMA(7, 1, 1) RMSE=1367.051
ARIMA(7, 1, 2) RMSE=1240.139
ARIMA(8, 0, 0) RMSE=1292.595
ARIMA(8, 0, 1) RMSE=1313.710
ARIMA(8, 0, 2) RMSE=1083.098
ARIMA(8, 1, 0) RMSE=1331.374
ARIMA(8, 1, 1) RMSE=1186.291
ARIMA(8, 1, 2) RMSE=1057.051
ARIMA(9, 0, 0) RMSE=1331.399
ARIMA(9, 0, 1) RMSE=1365.887
ARIMA(9, 0, 2) RMSE=1075.722
ARIMA(9, 1, 0) RMSE=1142.123
ARIMA(9, 1, 1) RMSE=1050.829
ARIMA(9, 1, 2) RMSE=1012.057
ARIMA(10, 0, 0) RMSE=1162.586
ARIMA(10, 0, 1) RMSE=1050.979
ARIMA(10, 0, 2) RMSE=1000.805
ARIMA(10, 1, 0) RMSE=956.875
ARIMA(10, 1, 1) RMSE=947.898
ARIMA(10, 1, 2) RMSE=957.215
ARIMA(11, 0, 0) RMSE=973.522
ARIMA(11, 0, 1) RMSE=939.855
ARIMA(11, 0, 2) RMSE=940.063
ARIMA(11, 1, 0) RMSE=927.846
ARIMA(11, 1, 1) RMSE=972.177
ARIMA(11, 1, 2) RMSE=916.859
ARIMA(12, 0, 0) RMSE=920.802
ARIMA(12, 0, 1) RMSE=924.431
ARIMA(12, 0, 2) RMSE=930.787
ARIMA(12, 1, 0) RMSE=977.595
ARIMA(12, 1, 1) RMSE=958.552
ARIMA(12, 1, 2) RMSE=955.259
Best ARIMA(11, 1, 2) RMSE=916.859
```

(a) Best ARIMA values

Dep. Variable:	y	**No. Observations:**	77
Model:	ARIMA(11, 1, 2)	**Log Likelihood**	-620.640
Date:	Sun, 01 May 2022	**AIC**	1269.280
Time:	20:45:29	**BIC**	1301.910
Sample:	0	**HQIC**	1282.321
	- 77		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5374	0.187	2.876	0.004	0.171	0.904
ar.L2	-0.7093	0.128	-5.537	0.000	-0.960	-0.458
ar.L3	-0.2741	0.193	-1.419	0.156	-0.653	0.104
ar.L4	-0.0133	0.194	-0.069	0.945	-0.394	0.367
ar.L5	-0.4821	0.156	-3.094	0.002	-0.787	-0.177
ar.L6	-0.0300	0.182	-0.165	0.869	-0.386	0.326
ar.L7	-0.2165	0.141	-1.532	0.125	-0.493	0.060
ar.L8	-0.2381	0.189	-1.259	0.208	-0.609	0.133
ar.L9	-0.1230	0.144	-0.856	0.392	-0.405	0.159
ar.L10	-0.2325	0.089	-2.626	0.009	-0.406	-0.059
ar.L11	0.1284	0.118	1.086	0.277	-0.103	0.360
ma.L1	-0.7864	0.166	-4.733	0.000	-1.112	-0.461
ma.L2	0.6324	0.107	5.922	0.000	0.423	0.842
sigma2	5.139e+05	9.91e+04	5.184	0.000	3.2e+05	7.08e+05

Ljung-Box (L1) (Q):	0.49	**Jarque-Bera (JB):**	14.43
Prob(Q):	0.48	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.99	**Skew:**	0.98
Prob(H) (two-sided):	0.97	**Kurtosis:**	3.83

(b) Model Summary

The Expected v/s Predicted graph is given as follows:

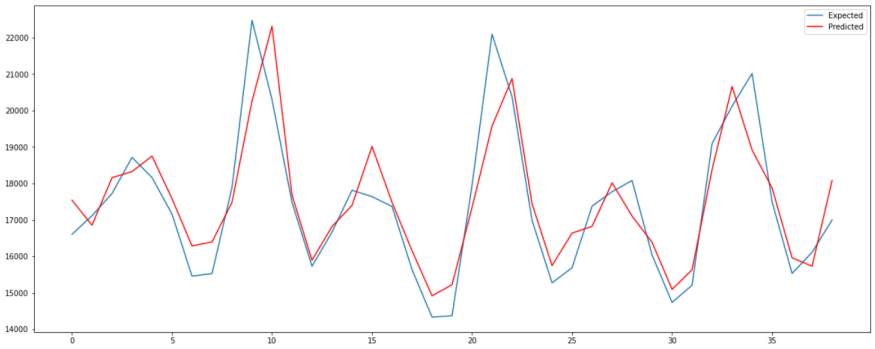

Figure 12: Expected V/s Predicted

5.2 Identifying Anomaly Dates

From the Expected v/s Predicted graph, it is observed that the demand in particualr months is lower than what was expected and in some months, it was higher. To investigate what the reason for these irregularities could be, here outlier dates are identified.

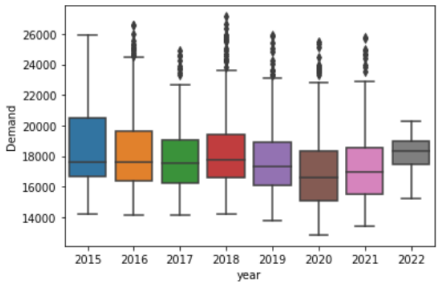

Figure 13: Boxplot of Demand v/s Year

The following is a box plot of the daily energy demand data. Boxplot gives a rough estimate of mean, median and range of demand values for all the years. From this, the outlier values are identified. The top quartile threshold is set at 95% and the bottom quartile threshold is set at 5%. All the dates beyond these thresholds are set aside as anomalies.

```
top_anomaly_dates
```

```
19      2015-07-20
27      2015-07-28
28      2015-07-29
29      2015-07-30
47      2015-08-17
           ...
2249    2021-08-27
2386    2022-01-11
2396    2022-01-21
2402    2022-01-27
2421    2022-02-15
Name: Date, Length: 128, dtype: object
```

```
bottom_anomaly_dates
```

```
95      2015-10-04
101     2015-10-10
102     2015-10-11
108     2015-10-17
116     2015-10-25
           ...
2314    2021-10-31
2376    2022-01-01
2377    2022-01-02
2418    2022-02-12
2439    2022-03-05
Name: Date, Length: 128, dtype: object
```

Figure 14: Anomaly Dates

5.3 Twitter Streaming

Tweets on the particular dates are streamed using snscrape library. As mentioned earlier, snscrape library allows streaming from old records of tweets. The tweets are saved as top_anomaly_twitter_data and bottom_anomaly_twitter_data.

5.3.1 Text Preprocessing

The Twitter data collected is not structured. To convert the data into a machine-readable format, a few text preprocessing steps need to be performed. The general steps followed are:

1) Removing stopwords, punctuation 2) Tokenizing words 3) Removing any noise (URLs, emojis, numbers etc.) 4) Lowercasing words 5) Lemmatizating text

The Visualization of dataframes before and aftertext preprocessing are shown in the following figures. The tweet length distribution is also plotted.

	Unnamed: 0	text	date	Tweet Id
0	0	If you live near New York's La Guardia Airport...	2015-10-04 23:54:01+00:00	650821202676719616
1	1	.@Space_Station Hello #ISS from New York City ...	2015-10-04 23:27:07+00:00	650814435032875008
2	2	@myitalianangels @joshcometomexic @joshgroban ...	2015-10-04 23:04:48+00:00	650808820478930944
3	3	really ⁉️, from 2014 🅰️ \n\n"Half Of New York Cit...	2015-10-04 22:58:17+00:00	650807179230998528
4	4	@JazzyBonesJones by train yeah. I live in NYS ...	2015-10-04 21:28:19+00:00	650784536855474176

(a) Before Pre-processing

	Unnamed: 0	text	date	Tweet Id
0	0	[live, la, guardia, airport, talk]	2015-10-04 23:54:01+00:00	650821202676719616
1	1	[hello, i, central, park, mi, away, issabove]	2015-10-04 23:27:07+00:00	650814435032875008
2	2	[myitalianangels, joshcometomexic, joshgroban,...	2015-10-04 23:04:48+00:00	650808820478930944
3	3	[half, living, poverty, cc, michaeljaco]	2015-10-04 22:58:17+00:00	650807179230998528
4	4	[jazzybonesjones, train, yeah, live, ny, live,...	2015-10-04 21:28:19+00:00	650784536855474176

(b) After Pre-processing

Figure 15: Dataframes before and After Text Preprocessing

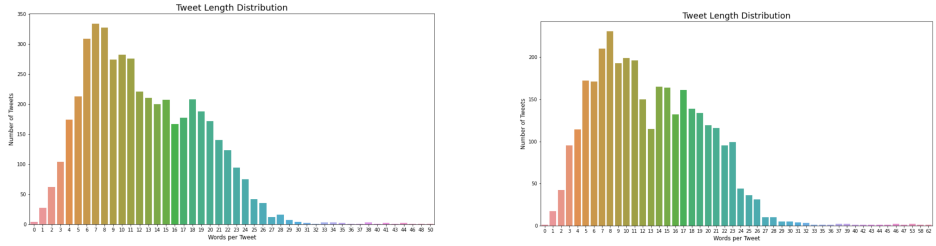

Figure 16: Tweet Length Distributions of Top and Bottom Dates

By creating a Bag of Words using Gensim's Dictionary constructor, each word in the tweets is paired with a unique integer identifier.

5.3.2 Topic Modeling

LDA or Latent Dirichlet Allocation is a popular Topic Modeling technique. It allows the grouping of unstructured data into a specific number of groups, thereby using unsupervised methods to cluster the data. LDA is applied to a document when the following conditions are assumed to The number of topics(k) is predetermined All other topic allocations for words except for the word at hand are true. The value of k is given as 5. This would group the words into sets of five topics.

The following are the pyLDA visualizations of top and bottom quartile dates.

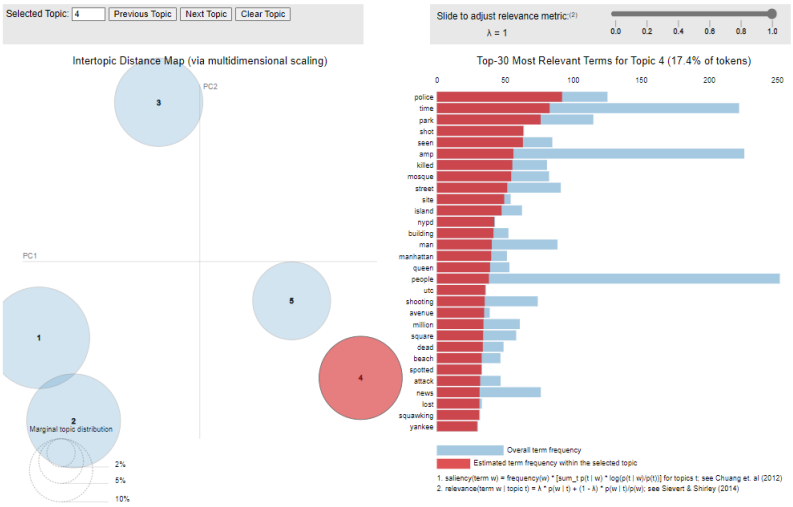

Figure 17: LDA of Top Extreme Values

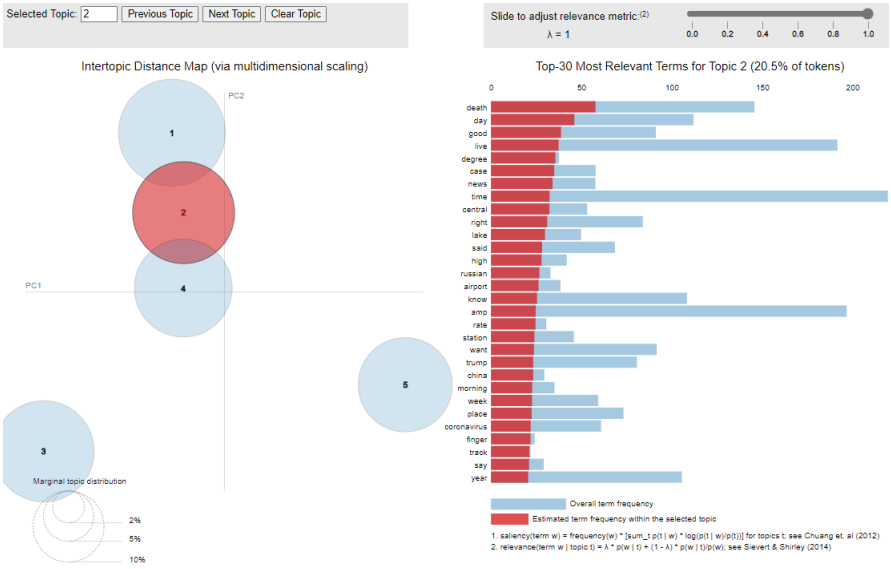

Figure 18: LDA of Bottom Extreme Values

LDA performs well on the data, however it is mostly designed for data written in paragraphs or larger texts. Since tweets are relatively small, GSDMM(Gibbs Sampling Dirichlet Multinomial Mixture) is used. GSDMM is also a topic modelling algorithm and is efficient on smaller textual data.

The following are the Wordcloud visualizations obtained on applying GSDMM on the data:

(a) Top Extreme Values **(b)** Bottom Extreme Values

Figure 19: GSDMM WordClouds

6 Results and Discussion

Using LDA, five topic clusters are generated for tweets on high(top) extreme and low(bottom) extreme dates.

For the Top extreme, the following are the possible contexts of the words

Topic 1 - The top five-word suggestion in this topic are positive and are indicated the future. This indicates that people could have been discussing innovations or the climate crisis.
Topic 2 - Words like a storm, lake etc indicate that these were rainy days with possible storms in New York.
Topic 3 - This cluster has negative words like death, train, military, hospital etc which could mean that there could have been an increase in criminal activity.
Topic 4 - This topic cluster also has negative words like kill, attack, mosque, manhattan, shooting etc which also indicates high criminal activity.
Topic 5 - This topic cluster is a mixture of words like weather, peak, wind etc which means that people were discussing changing weather patterns. Words about criminal activity are also seen.

For the lower extreme, the following are the possible contexts of the words

Topic 1 - This topic cluster has words like love, border, country which are indicative of a feeling of patriotism within the general populace
Topic 2 - Words like Coronavirus, china, Russia, death etc are highlighted in this topic which could mean an increase in the number of Covid cases around the world.
Topic 3 - Protest, government, death, police etc are the words highlighted in this cluster. This means there were some negative sentiments associated with government regulation on these days.
Topic 4 - This cluster is a mixture of positive and negative words with ambiguous meanings.
Topic 5 - Trump, news etc are highlighted pointing to some government activity. Words like game, future, and hot are also present in this cluster. This could mean a change in weather.

GSDMM gives more insight into the topics. A total of twelve topic clusters are plotted by GSDMM.

For the Top extreme, the following are the possible contexts:

Topic 1 - People are discussing the future year (positive)
Topic 2 - Discussions on rivers, islands, hotels (positive)
Topic 3 - Discussion on killing, shooting in mosques, Brooklyn (negative)
Topic 4 - Discussion on killing, police (negative)
Topic 5 - Discussion on events/concerts that are coming soon (positive)
Topic 6 - Discussion on weather conditions mostly heat and highs
Topic 7 - Discussion on some criminal activity (negative)
Topic 8 - Discussion about Trump
Topic 9 - Discussion about the military, border
Topic 10 - Discussion on heat/weather
Topic 11 - Discussion about sharks, islands, attacks (negative)
Topic 12 - Unclear discussion

For the Bottom extreme, the following are the possible contexts:

Topic 1 - Discussion about Trump
Topic 2 - Discussion about life(positive)
Topic 3 - Discussion about Coronavirus and death(negative)
Topic 4 - Discussion about some games(positive)
Topic 5 - Discussion about someone missing, police (negative)
Topic 6 - Criminal activity discussion(negative)
Topic 7 - Unclear
Topic 8 - Wine, orchards(positive)
Topic 9 - fire, chemical (negative)
Topic 10 - Undersea concerns with Russia
Topic 11 - Discussion about Technology
Topic 12 - lake, good, rose etc(positive)

Through visual inspection, it is seen that there are more negative discussions than positive ones. There seems to be a lot of discussion on some criminal activity in the city on days of high extremes. A significant amount of discussion is revolved around weather conditions which means that there is some truth in the energy demand prices being associated with weather. Topics of those pertaining to the government and military also occur in the topic clusters, the nature of which are ambiguous. Words related to technology and future are also witnessed.

7 Conclusion

Energy Demand Forecasting would enable suppliers to allocate adequate resources for electricity generation. It would allow residents to plan their household budget in accordance with the fluctuating demand for electricity. In this project, ARIMA, a seasoned tool for Time Series Forecasting, is used to forecast the energy demand in the state of New York. This project also investigates the reason behind the occurrence of anomalies; very high or very low demand for electricity. Twitter Data is streamed for the anomalous dates and Topic Modelling techniques like LDA and GSDMM are used to conduct analysis. On studying the tweets, it is found changing weather conditions may not be the only reason behind the spikes in demand. It also appears that people are discussing some criminal activity on days of spikes in demand. Possible connections to government and technology are also observed.

8 Future Scope

From analysis, increased criminal activity was observed to be a possible reason for the increase in energy demand. This could mean that the increase in criminal activity may have a causal relationship with the spikes in energy demand. Further exploration on the causality link can be studied. Investigation on whether these patterns are specific to anomalous dates can also be conducted.

References

[1] Mellow, Sophie. "New York's eye-watering energy price hikes hit home as 1.3 million residents fall behind on bill payments." Fortune, Fortune, 18 Mar.2022,https://fortune.com/2022/03/18/new-york-energy-price-hikes-con-edison-million-residents-behind-on-bill-payments/

[2] I. A. S. Abu Amra and A. Y. A. Maghari, "Forecasting Groundwater Production and Rain Amounts Using ARIMA-Hybrid ARIMA: Case Study of Deir El-Balah City in GAZA," 2018 International Conference on Promising Electronic Technologies (ICPET), 2018, pp. 135-140, doi: 10.1109/ICPET.2018.00031.

[3] Y. Wang and Y. Guo, "Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost," in China Communications, vol. 17, no. 3, pp. 205-221, March 2020, doi: 10.23919/JCC.2020.03.017.

[4] A. Gupta and A. Kumar, "Mid Term Daily Load Forecasting using ARIMA, Wavelet-ARIMA and Machine Learning," 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe), 2020, pp. 1-5, doi: 10.1109/EEEIC/ICPSEurope49358.2020.9160563.

[5] Y. Du, "Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network," 2018 Chinese Control And Decision Conference (CCDC), 2018, pp. 2854-2857, doi: 10.1109/CCDC.2018.8407611.

[6] A. MITKOV, N. NOORZAD, K. GABROVSKA-EVSTATIEVA and N. MIHAILOV, "Forecasting the Energy Consumption in Afghanistan with the ARIMA Model," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), 2019, pp. 1-4, doi: 10.1109/ELMA.2019.8771680

[7] F. Alves, P. M. Ferreira and A. Bessani, "Design of a Classification Model for a Twitter-Based Streaming Threat Monitor," 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2019, pp. 9-14, doi: 10.1109/DSN-W.2019.00010

[8] K. Okamoto and K. Yanai, "Analyzing Regional Food Trends with Geo-tagged Twitter Food Photos," 2019 International Conference on Content-Based Multimedia Indexing (CBMI), 2019, pp. 1-4, doi: 10.1109/CBMI.2019.8877473.

[9] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar and R. Pranata, "Twitter Topic Modeling on Football News," 2018 3rd International Conference on Computer and Communication Systems (ICCCS), 2018, pp. 467-471, doi: 10.1109/CCOMS.2018.8463231

[10] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). Association for Computing Machinery, New York, NY, USA, 233–242. https://doi.org/10.1145/2623330.2623715