The Alleys of American Economy

Debodeep Banerjee (1901253)
Sapienza University of Rome

Contents

Webscrapping

- All the links were scrapped from the following parent link.

`https://www.bis.org`

- Only speeches of USA and China have been collected.
- **Difficulties:** The links were divided into html and pdf. Hence few links were dropped.
- After scrapping, the texts of each link were stored as separate .txt file with which a parent dataframe was constructed where eacg row correspons to each text file.
- After this point, we move to data pre-processing.

Data pre-processing

Pre-processing

- removing stopwords
- removing punctuations and other symbols
- Lemmatization
- Process
  - First we remove all the stopwords and symbols using nltk library. The respective column shown in the figure is *cleaned speech*
  - After that we lemmatize all the columns using gensim library.The final column we obtained after pre-processing is 'final cleaned speech' which is shown in the picture.

	speeches	cleaned_speech	final_cleaned_speech
0	As prepared for deliveryGood afternoon, and we...	prepared deliverygood afternoon welcome im joh...	prepared deliverygood afternoon welcome im joh...
1	IntroductionThank you to the American Bankers ...	introductionthank american bankers association...	introductionthank american bankers association...
2	As prepared for deliveryGood afternoon, everyo...	prepared deliverygood afternoon thank st lawre...	prepared deliverygood afternoon thank st lawre...
3	Seventeen months have passed since the U.S. ec...	seventeen months passed economy faced force co...	seventeen months passed economy faced force co...
4	Thank you, Gigi and Scott, and thanks to every...	thank gigi scott thanks work bringing today pa...	thank gigi scott thanks work bringing today pa...
...	...	...	...
682	It is a great pleasure to have the opportunity...	great pleasure opportunity speak today remarks...	great pleasure opportunity speak today remarks...
683	It is a pleasure to offer a few remarks at thi...	pleasure offer remarks conference marking th a...	pleasure offer remarks conference marking th a...
684	The thoughts that follow are my own, and are n...	thoughts follow necessarily shared colleagues ...	thoughts follow necessarily shared colleagues ...
685	Good afternoon. I am delighted to have the opp...	good afternoon delighted opportunity participa...	good afternoon delighted opportunity participa...
686	Let me start by thanking the organizers for in...	let start thanking organizers including event ...	let start thanking organizers including event ...

687 rows × 3 columns

Exploratory Data analysis: Strategy

In this part, we aim to explore the texts that we have obtained. Our main motto is to find out the words with the highest and lowest frequencies respectively. In order to obtain that, we maintain the following procedures.

- **Capture all the words**

In this part, as we have to conduct some sort of histogram analyis, we contain all the texts from all the rows in a single basket (technically, a list). The purpose of doing this is described below.

- **Tokenizing**

Once we contain all the text in a single basket, the next task is to express each word as a separate *entity*. We can do this task by tokenizing the basket that we have obtained before. Once we express each word as a dedicated object, there will be no problem in calculating the frequencies of the words.

Exploratory Data analysis: Strategy

• After tokenizing, the frequencies of each words were counted and stored in a python dictionary. In this case, the words are the keys and corresponding frequencies as the values.

• **Preliminary observation**

- Highest frequency: **'Financial'**
- Lowest frequency: **'yorkim'**

• Our next goal is to obtain a graphical representation of top 5 and top 10 highest used words.

Exploratory Data analysis: Visualization

- **Top 5 highest used word**

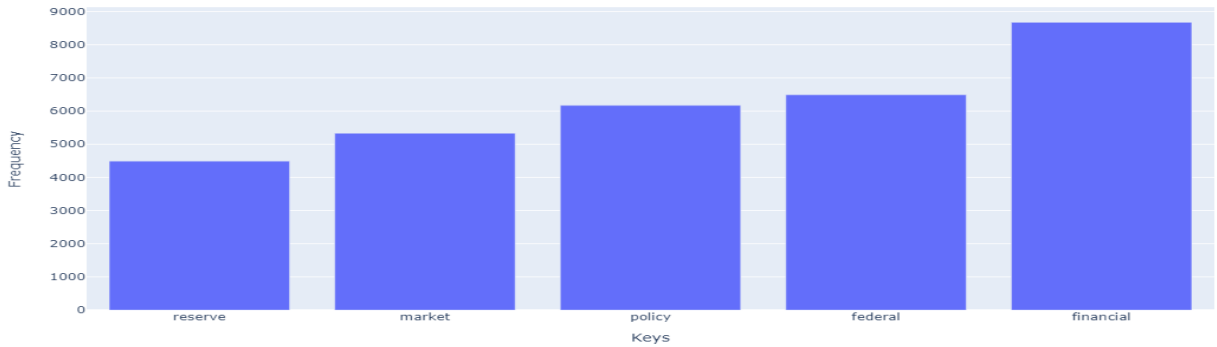
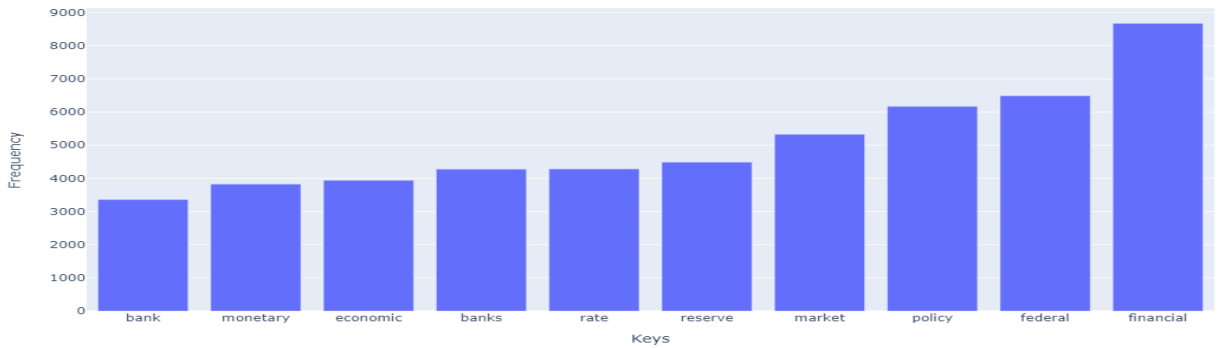


Figure: Top 5 highest used word

- **Top 10 highest used word**



Exploratory Data analysis: Visualization

Word cloud

Figure: Worldcloud

Latent dirichlet allocation

• The latent dirichlet allocation(LDA) is essentially used to identify the relevant topics and corresponding most important words in a document.
• In order to conduct an LDA, we have to tokenize each document, i.e. each row of the data. After that we remove those words which are used in less than 15 documents. The final choice we have to make before fitting the model is the choice of number of topics. In order to do that, we seek help of the coherence score of the model we fit. Here we set a list of numbers and based on those, we see the coherence score. Eventually we pick the number corresponding to the highest coherence score.

Latent dirichlet analysis

- Number of topics: 27
- Coherence score: 0.2682616483498412

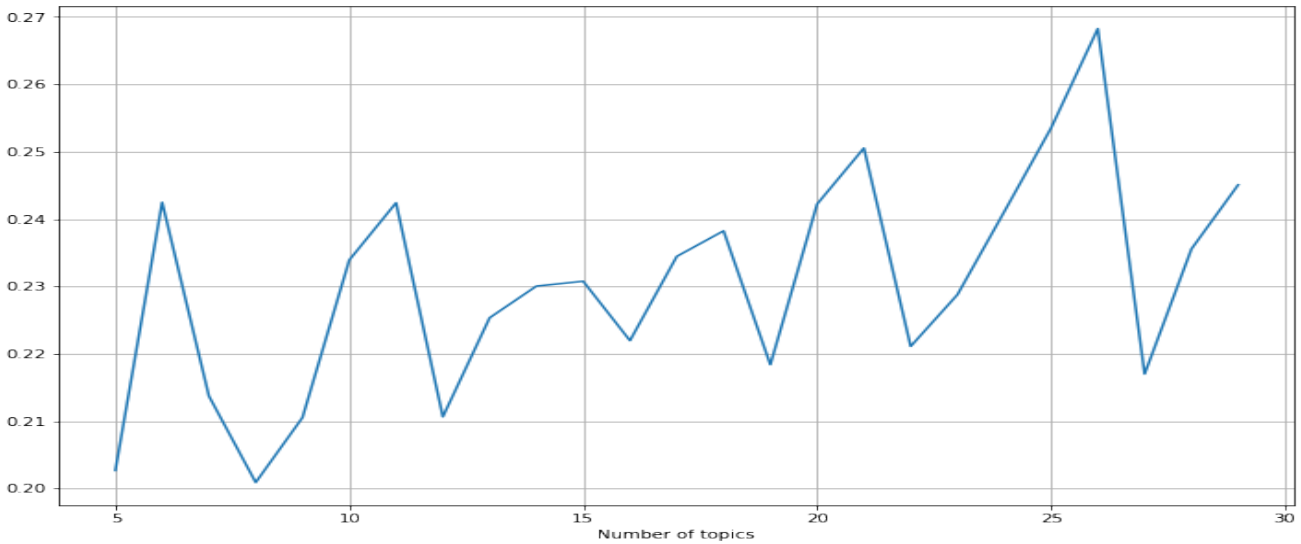


Figure: Plot of cohrence scores

Latent dirichlet allocation

• After completing the LDA implementation, we find the following outcome. The link of the interactive plot is here

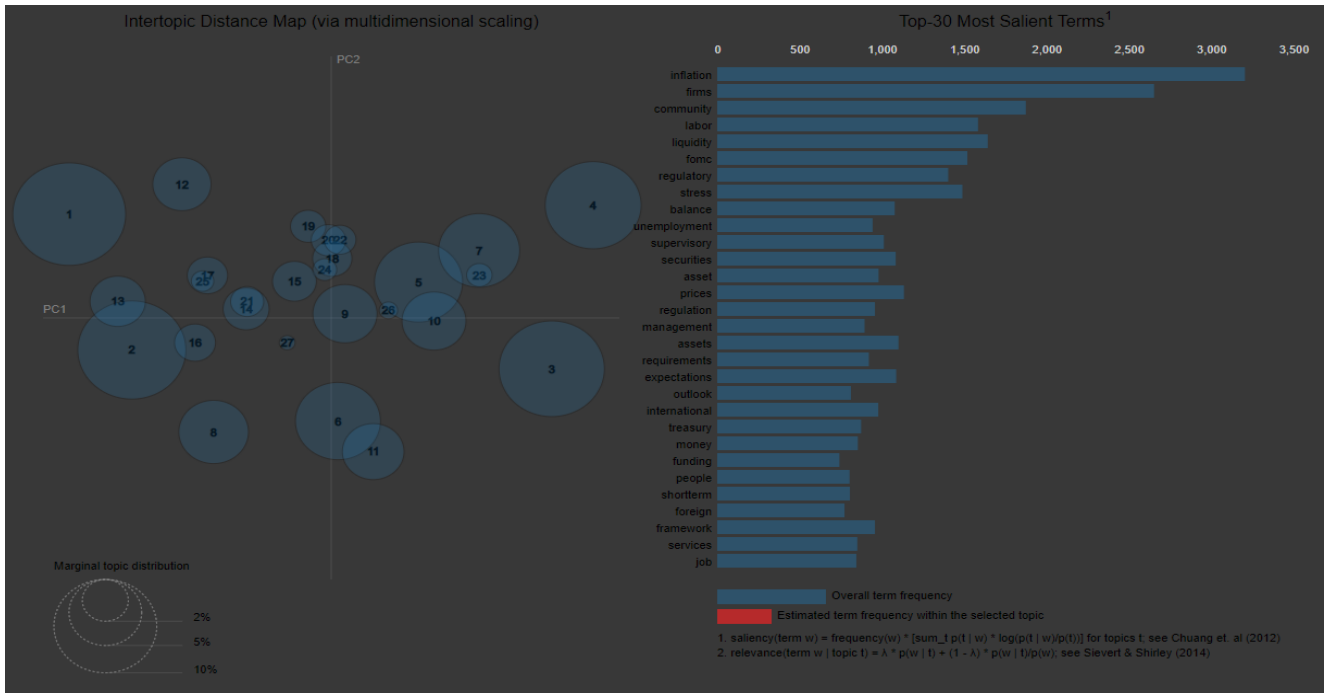


Figure: Latent dirichlet allocation

Forecasting: Data collection and pre-processing

- **The Data**

The data was collected from `https://fred.stlouisfed.org/`. We have collected 10 years of daily inflation data. The aim is to find a prediction model.

- Pre-processing
  - The data contained some missing plots filled with '.' values. So had to omit those cells before proceeding.
  - The numeric column was string in nature. So we made it as a float data.

Exploratory data analysis

The plot below is representing the inflation rate of the USA from 01/02/2008 to 17/09/2021.

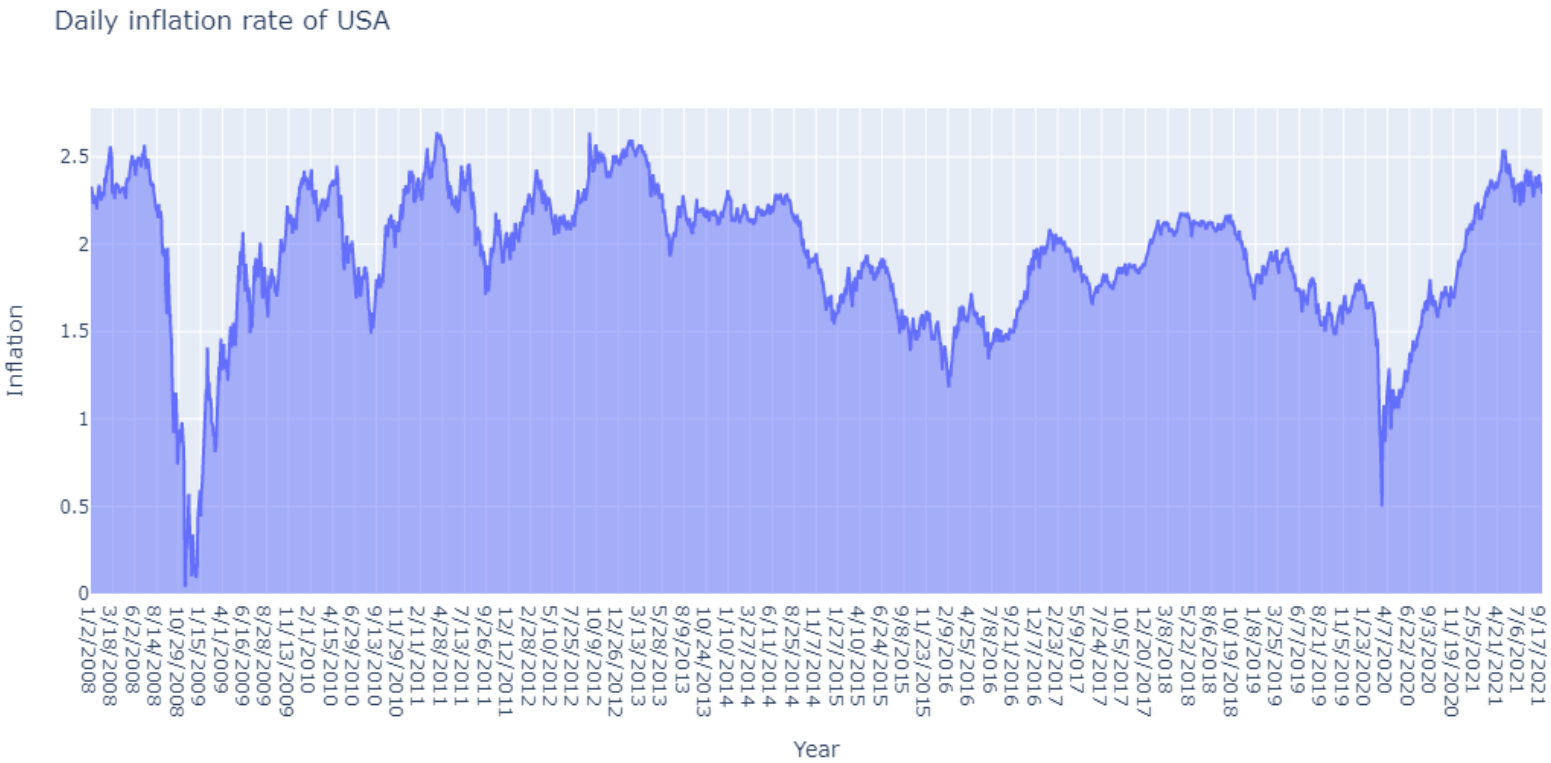
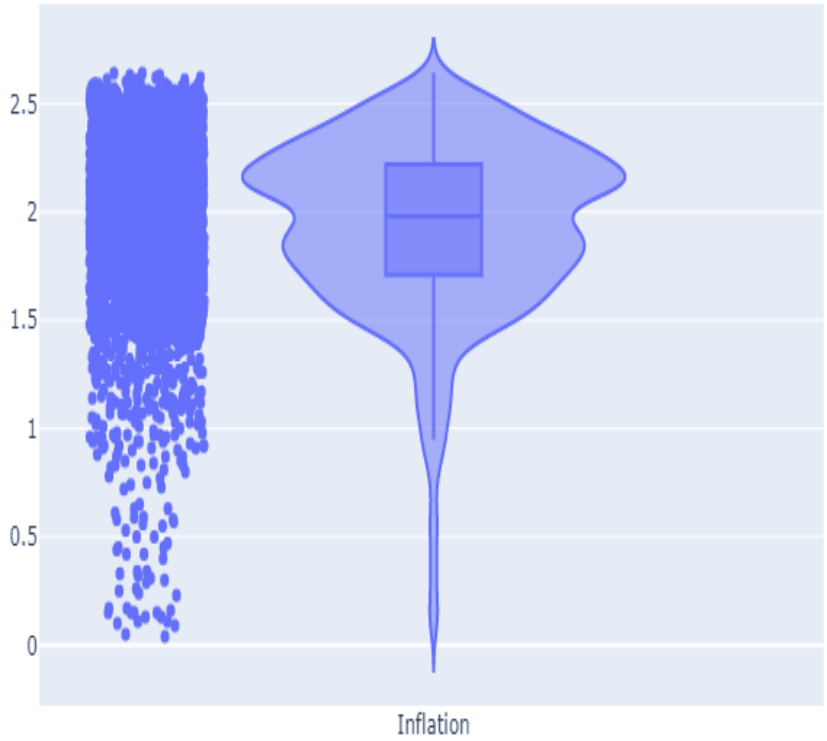


Figure: Inflation rate of the USA

Exploratory data analysis: Violin plot



Violin plot

- Maximum value: 2.64
- Minimum value: 0.04
- Median value: 1.98
- 1st quartile: 1.71
- 3rd quartile: 2.32

Exploratory data analysis: Rolling and Exponential mean

- 12 days moving and exponential average

Figure: 12 days moving and exponential average

Exploratory data analysis: Rolling and Exponential mean

- 30 days moving and exponential average

Figure: 30 days moving and exponential average

Exploratory data analysis: Rolling and Exponential mean

- 60 days moving and exponential average

Figure: 60 days moving and exponential average

Exploratory data analysis: Rolling and Exponential mean

- 90 days moving and exponential average

Figure: 90 days moving and exponential average

Exploratory data analysis:Trend, Seasonality

The following image examines mainly the nature of the trend and seasonality of the data. We can observe some sort of seasonality in the data.

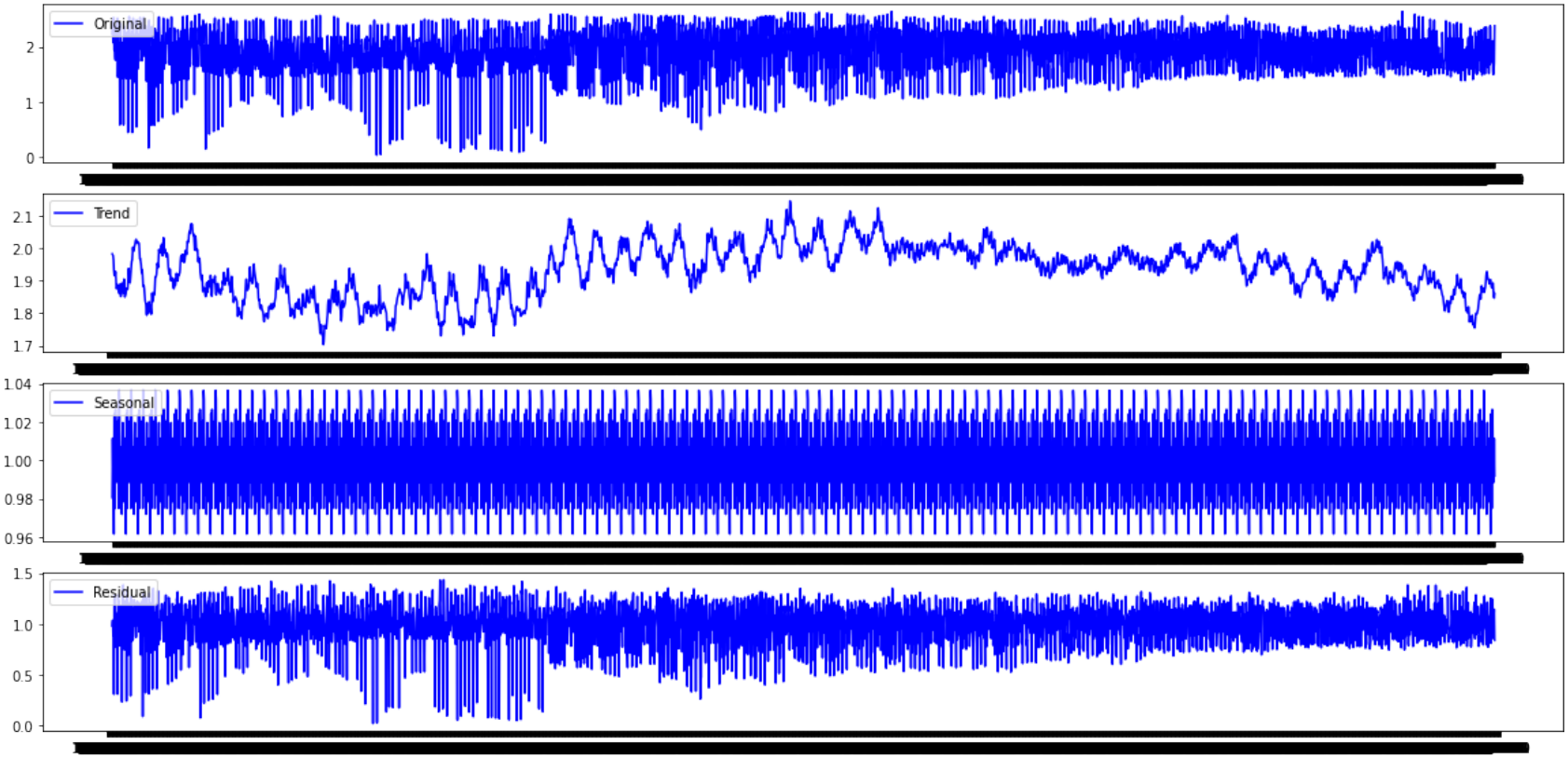


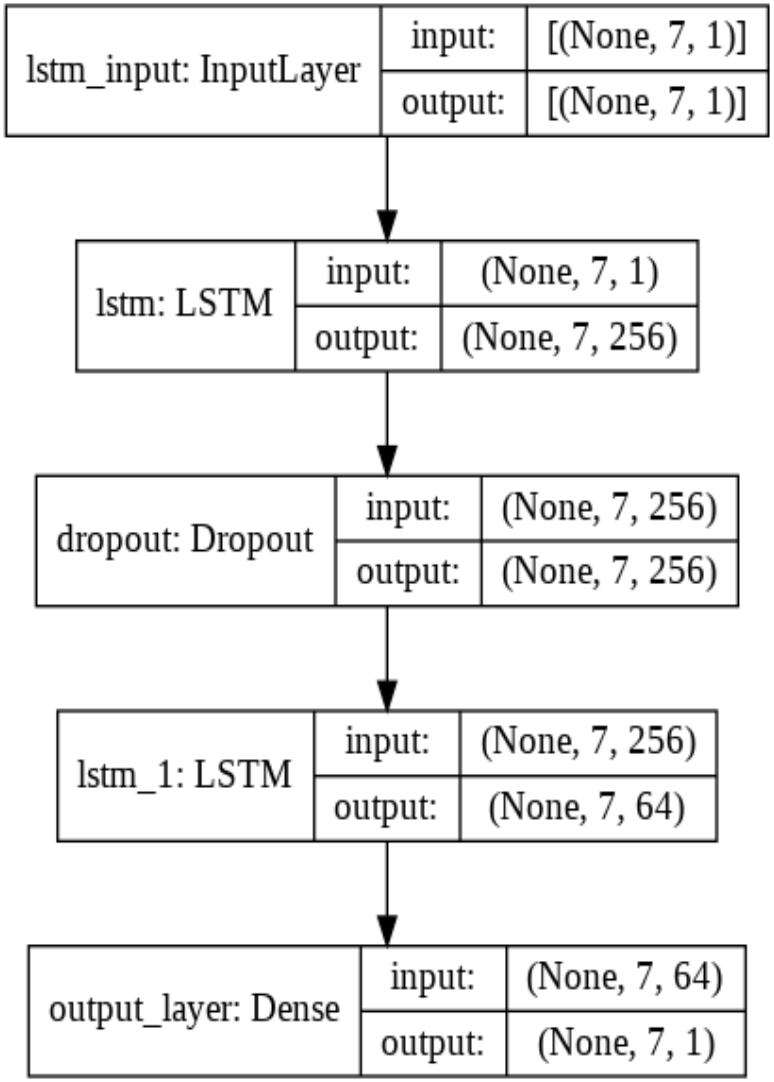
Figure: Trend, seasonality and residual check

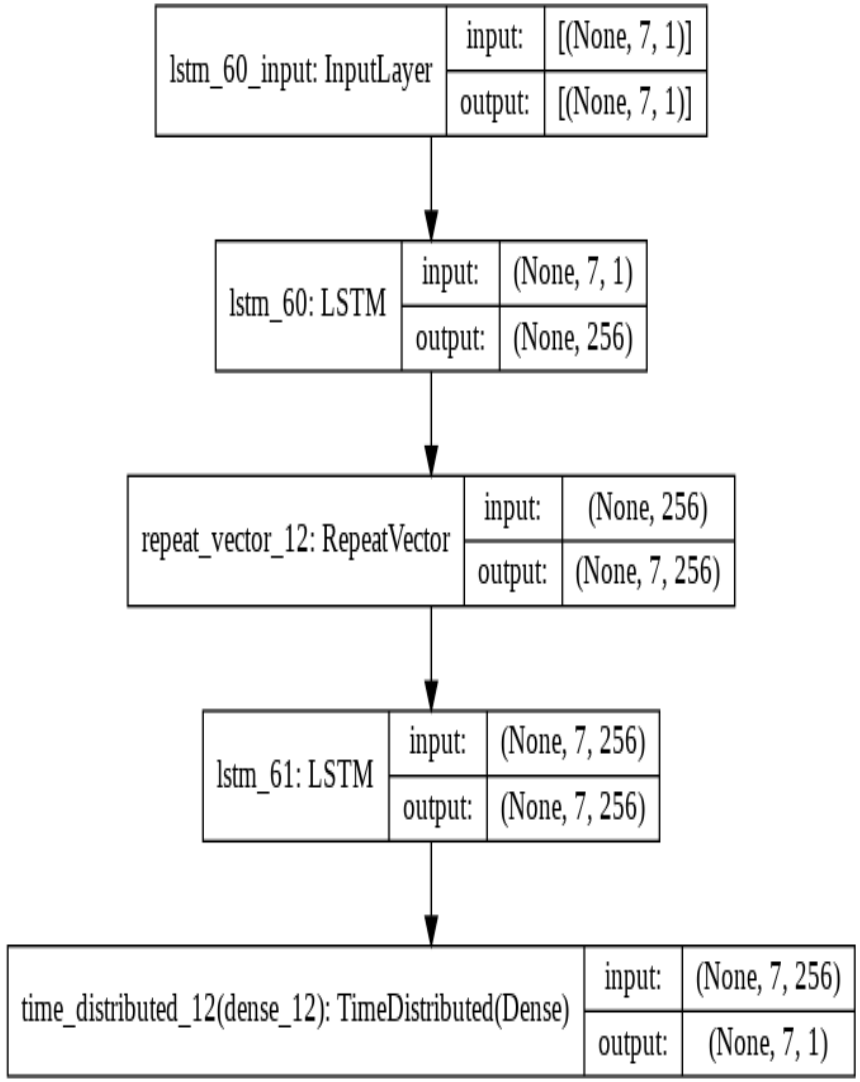
Forecasting: LSTM and LSTM Autoencoder

• In this part, we try to set up a model which can be used to predict the inflation rate in future. In order to achieve that, we choose to work with two deep learning models, viz. LSTM and LSTM autoencoder. The difference between LSTM and LSTM auto encoder is that, the auto enocoder, instead of directly generating the outcome, it first encodes the data and with a dedicated decoder function, the output is generated. In many cases, the LSTM AE works better than the classical LSTM set up.

• As the LSTM doesn't use the entire past data and use a subset of it, we have to specify a time window with which it determines how much data to remember. In our study, we choose to keep a one week time window.

Forecasting: Model architectures





Forecasting: Model Evaluation

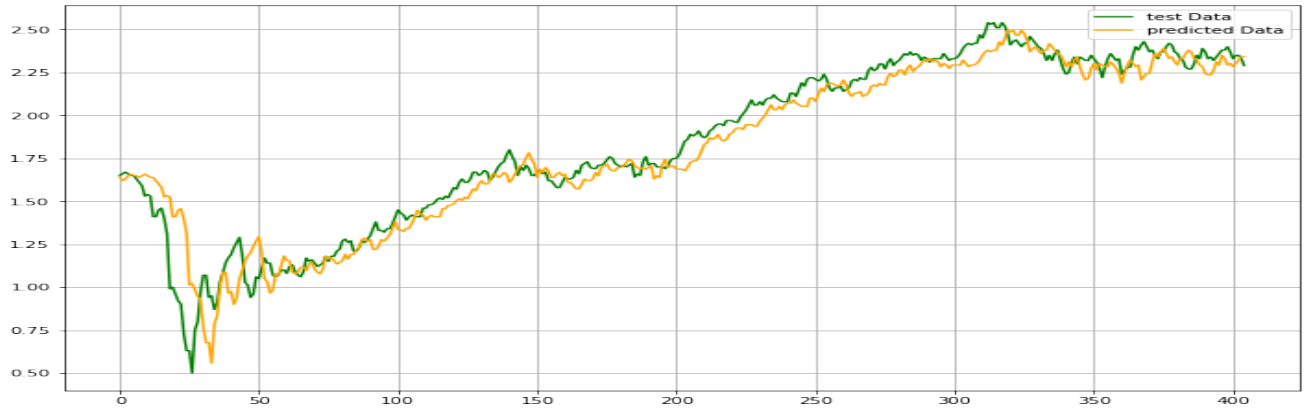


Figure: RMSE=0.123

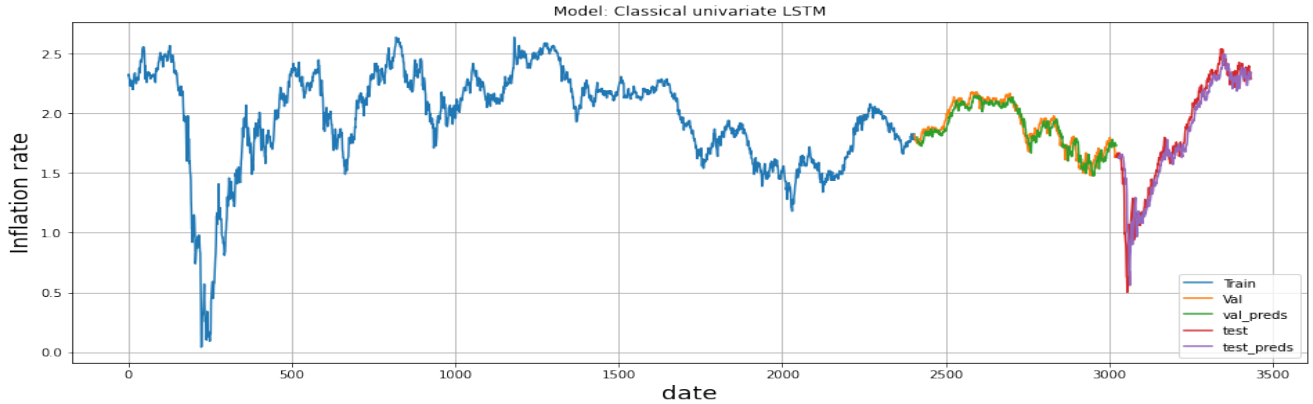


Figure: LSTM- Train, valid, test and predicted

Forecasting: Model Evaluation

Figure: RMSE= 0.093

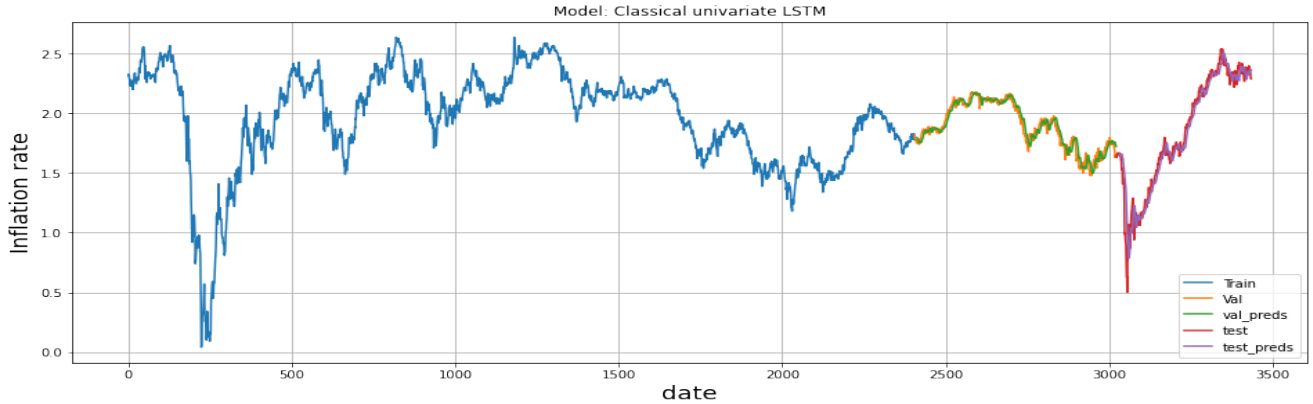


Figure: LSTM auto encoders- Train, valid, test and predicted

Discussion and conclusion

• As we have taken a combination of USA and China speeches, and considering the fact that they are two biggest competing economic force, 'inflation' and 'Unemployment' are two words that are used in almost all the topics. It it bit intuitive but we have the opportunity to claim that inflation and unemployment are something to be concerned of.

• A date wise collection of textual data may help us to formulate a model that'd help us in predicting the inflation rates. Probably newspaper data or social media data is more suitable in predicting using textual data as there will be no missing dates. In our case we have 1212 files and more than 3000 dates for inflation rates.

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
2. Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 57, 102282.
3. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2