SVKM'S NMIMS Deemed to be UNIVERSITY

NILKAMAL SCHOOL OF MATHEMATICS,
APPLIED STATISTICS & ANALYTICS

TEXT-BASED COAL PRICE FORECASTING USING LSTM & NLP TECHNIQUES

Team Members:

Salita D'britto (A010)

Jay Kangane (A028)

Advait Raut (A053)

Indrayani Shinde (A065)

Akash Yadav (A072)

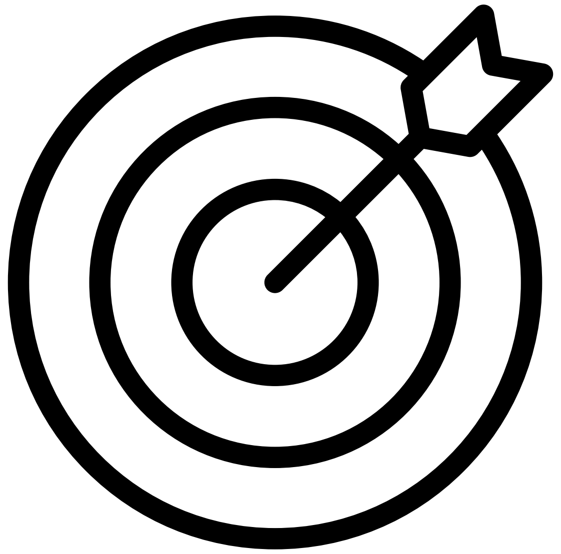
Mentor:

Mr. Rutik Bhurke

INTRODUCTION

- For the purpose of giving government and policymakers critical information and early warnings, an accurate forecasting model for future variations in coal prices is essential.
- First off, knowing what the trends in energy prices will be in near future helps producers determine whether an investment in the energy sector is sustainable. In contrast, consumers gain from projected energy prices by assessing how affordable energy will be in the future.
- Although price forecasting can be done based on financial indices, news articles can have a major influence on coal prices
- NLP allows us to extract valuable information from text data. Such information can be useful for predictive modelling, and can add additional features to the model.

OBJECTIVES



- Gather news article headlines, Coal Price, financial indices from **1st January, 2017** to **4th April, 2024**.
- Use LDA to get average daily sentiment scores.
- Fit models like ARIMAX, Random Forest and LSTM for Prediction.
- To check if incorporating News Headlines' Sentiment Scores improves the models' evaluation.

LITERATURE REVIEW

Impact of News on the Commodity Market: Dataset and Results

Ankur Sinha, Tanmay Khandait

Production and Quantitative Methods
Indian Institute of Management Ahmedabad
Ahmedabad, India 380015
asinha@iima.ac.in, tanmayk@iima.ac.in

This research paper proposes a framework for extracting various dimensions of information from news headlines related to the gold commodity.

LITERATURE REVIEW

Text-based crude oil price forecasting: A deep learning approach

Xuerong Li [a], Wei Shang [b,a,*], Shouyang Wang [a,b]

[a] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[b] Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China

This paper emphasizes the importance of combining text features and financial features for improved forecasting accuracy.

DATA PREPARATION

- **NEWS HEADLINES**

Daily news headlines and corresponding dates related to **Coal, Energy, Oil & Natural gas** are collected from various authentic sources like worldcoal.com, investing.com, reuters.com, etc.

Time period: 1st January 2017 to 4th April 2024

- **COAL PRICE**

Daily Coal price of **United States of America** is collected from Businessinsider.com

DATA PREPARATION

- **FINANCIAL INDICES**

The Financial Indices are fetched from authentic sources like investing.com. The following financial indices are collected:-

- **Richards Bay Coal Futures**
- **Newcastle Coal Futures**
- **Argus-McCloskey Coal Futures**
- **Bloomberg Commodity Index**
- **Dow Jones Commodity Index**

DATA PREPARATION

- **SENTIMENT SCORES**

RoBERTa is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. It is used to get sentiment scores of each news headline.

The following sentiment scores are calculated:

1. Polarity (-1 to 1): 1 represents positive sentiment and -1 represents negative sentiment
2. Subjectivity (0 to 1): 1 indicates personal opinion and 0 indicates factual information

DATA PREPARATION

- **SENTIMENT SCORES - POLARITY**

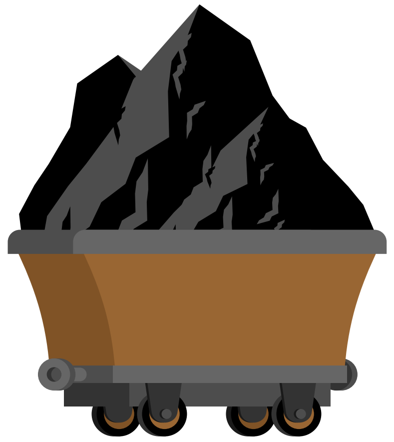
HEADLINE	POLARITY
Explosion in Pakistan coal mine kills 12 miners	**-0.9231**
Lithium on its way up, hopeful juniors say	**0.8725**

DATA PREPARATION

- **SENTIMENT SCORES - SUBJECTIVITY**

HEADLINE	SUBJECTIVITY
Plunging Natural Gas Prices Is Bad News for Drillers	**0.9187**
Russia’s Oil Exports By Sea Hit New 2024 Record	**0.1877**

DATA PRE PROCESSING



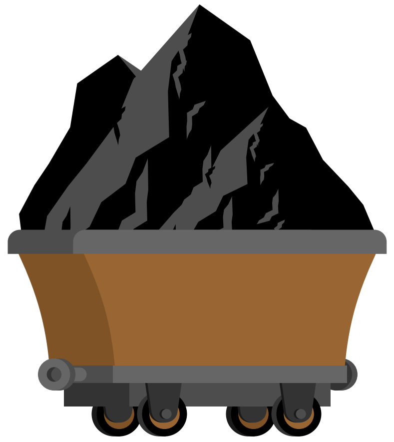
- **TEXT DATA**

Text pre-processing involves transforming raw text into a more suitable format for analysis and modelling.

Preprocessing steps:

Tokenization → **Stopwords Removal** → **Lemmatization** → **Remove Rare words** → **Remove frequent words**

DATA PRE PROCESSING



Original Sentence:
Tata Power to offset losses due to higher coal prices.

Tokenization: [’tata’, ‘power’, ‘to’, ‘offset’, ‘losses’, ‘due’, ‘to’, ‘higher’, ‘coal’, ‘prices’, ‘.’]

Removing stopwords: [’tata’ , ‘power’ , ‘offset’ , ‘losses’ , ‘higher’ , ‘coal’ , ‘prices’]

Lemmatization: [‘tata’ , ‘power’ , ‘offset’ , ‘loss’ , ‘high’ , ‘coal’ , ‘price’]

Tokenization → **Stopwords Removal** → **Lemmatization** → **Remove Rare words** → **Remove frequent words**

DATA PRE PROCESSING

Rare Words Removal:

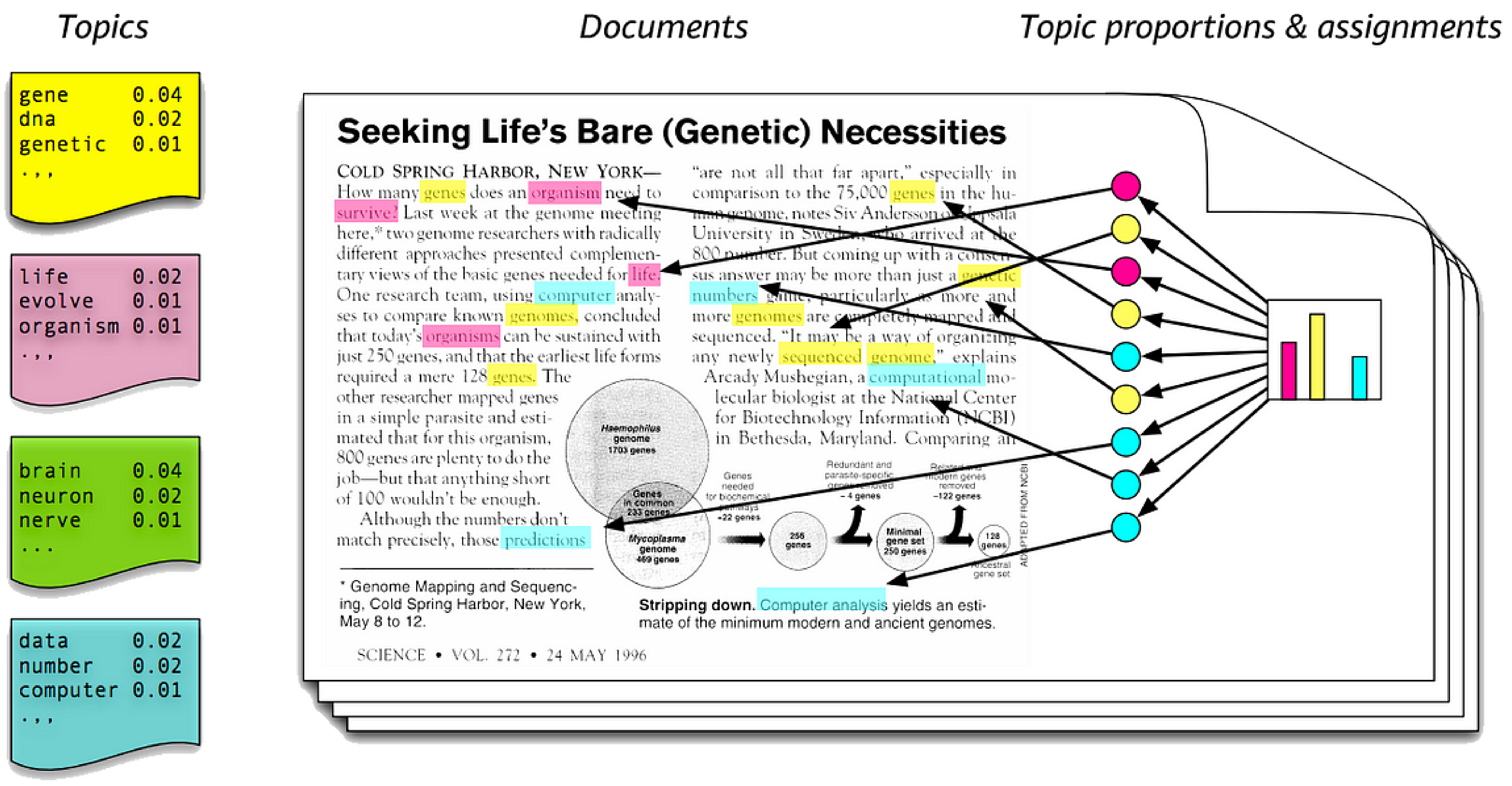
Words like **fast, expert, bounce** having word frequency **below 15** are removed from the corpus because they rarely offer anything interesting.

Frequent Words Removal:

Words like **Coal, Price** having word frequency **above 1000** are removed to reduce noise and improve models ability to identify meaningful patterns.

TOPIC MODELLING

Topic Modelling is a method for ***unsupervised*** classification of documents.



The main purpose of this statistical modelling algorithm is to understand the topics in the input data.

TOPIC MODELLING

Topic Distribution

Document 1
ball
ball
ball
planet
galaxy

Document 2
law
planet
law
law
law

Document 3
planet
planet
galaxy
planet
ball

Document 4
planet
galaxy
ball
planet
ball

Document	Sports	Science	Politics
1	0.6	0.4	0
2	0	0.2	0.8
3	0.2	0.8	0
4	0.4	0.6	0

TOPIC MODELLING

Word Distribution:

Document 1
ball
ball
ball
planet
galaxy

Document 2
law
planet
law
law
law

Document 3
planet
planet
galaxy
planet
ball

Document 4
planet
galaxy
ball
planet
ball

Topic	Ball	Law	Planet	Galaxy
Science	0	0	0.7	0.3
Sports	1	0	0	0
Politics	0	1	0	0

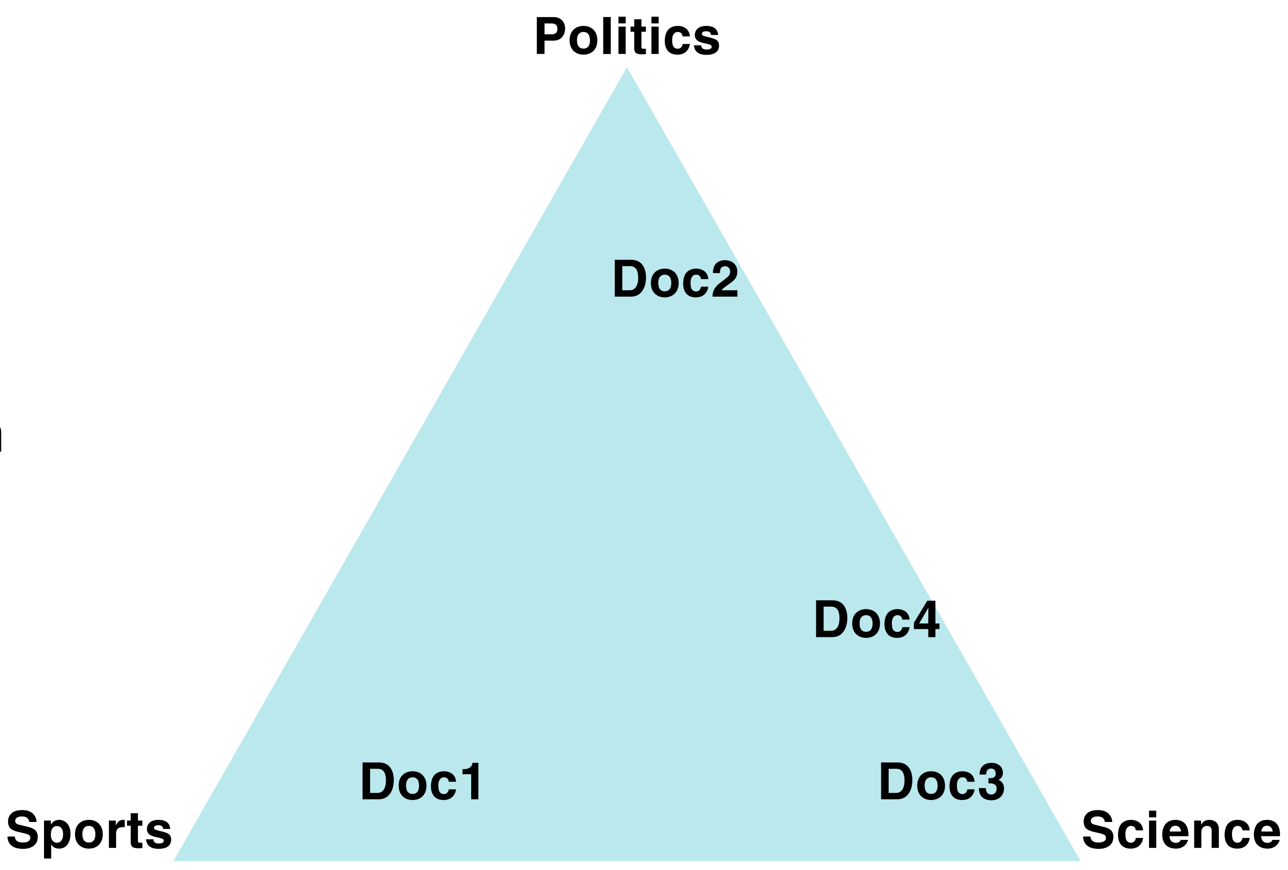
TOPIC MODELLING

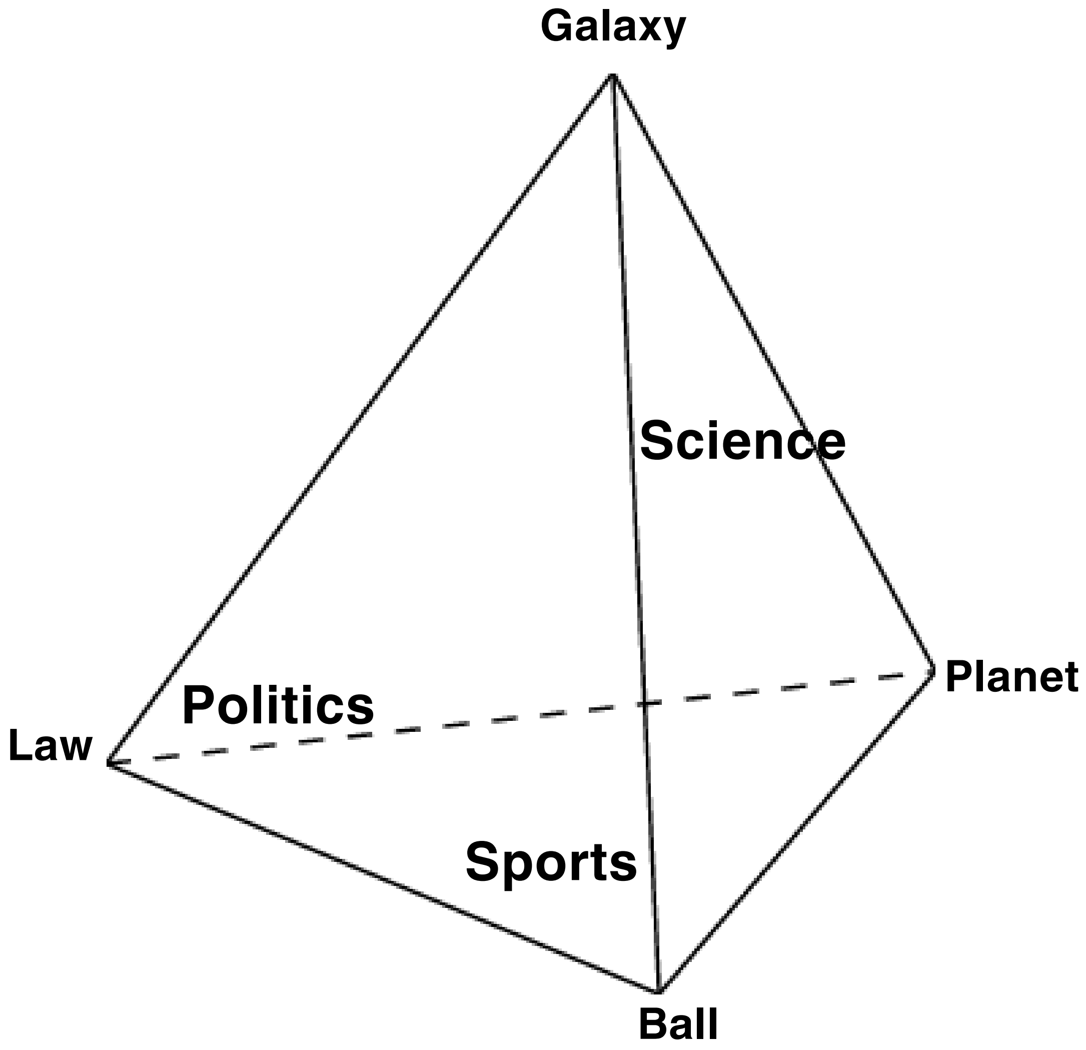
Latent Dirichlet Allocation has two components:

1. Distribution of topics in a document
2. Distribution of words in a topic

Thus, we shall get 2 sets of distributions:

Topic Distribution





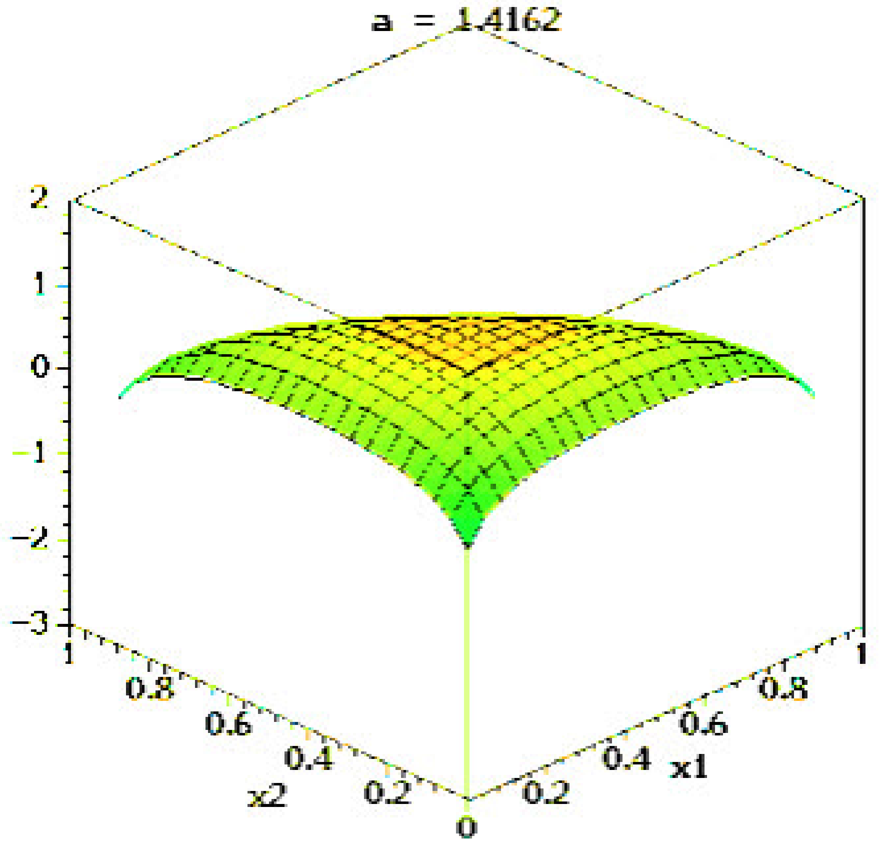
Word Distribution

TOPIC MODELLING

Probability that a word in a document is associated with topic j can be expressed as follows:

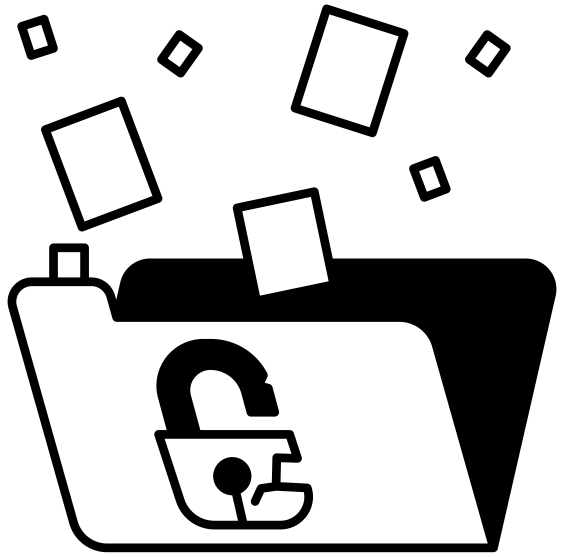
$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

Alpha and **eta** are hyperparameters, coming from Dirichlet distributions used to model the distribution of topics within documents and the distribution of words within topics.



PDF of Dirichlet Distribution when K = 3 as we change the vector α from α = (0.3, 0.3, 0.3) to (2.0, 2.0, 2.0)

TOPIC MODELLING

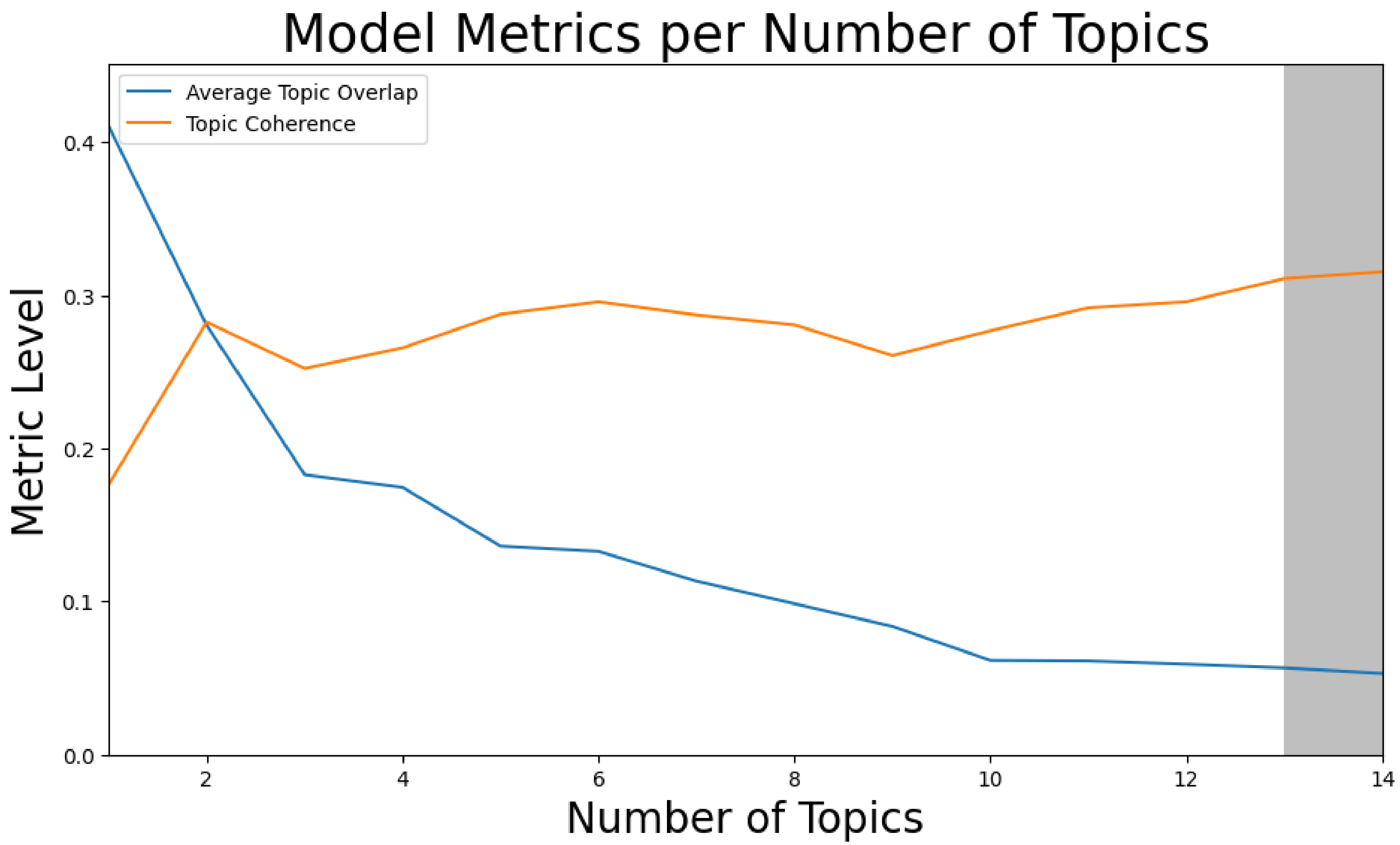


Selecting Optimum Value of K (Number of topics)

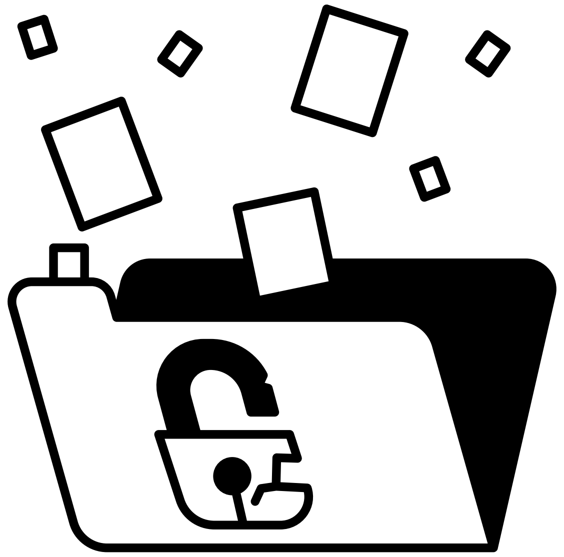
Coherence: measures quality of topics generated. Higher values indicate that intra-topic similarity is high.

Overlap: measures the inter-topic similarity. Higher value indicates that inter-topic similarity is high.

Optimum value of K: 6 Topics



TOPIC MODELLING



Topic Names

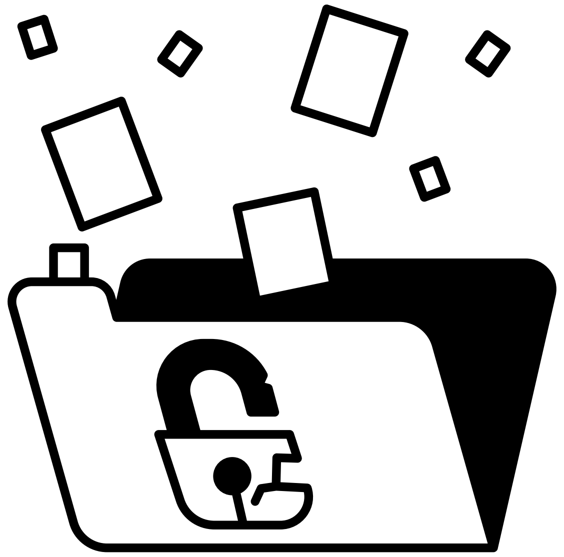
Based on the top 15 words in each topic, the topics are given names. Example:

Most frequent words of topic 2: russia, iran, opec, russian, sanction, india, deal, product, export, venezuela, boost, south, tanker, million, iranian

Topic Name: International Relations and Trade

Energy Markets & Operations	**International Relations & Trade**	**Environmental Sustainability**
Saudi Arabia & Investment	**China & Global Demand**	**Crude Oil Market Analysis**

TOPIC MODELLING



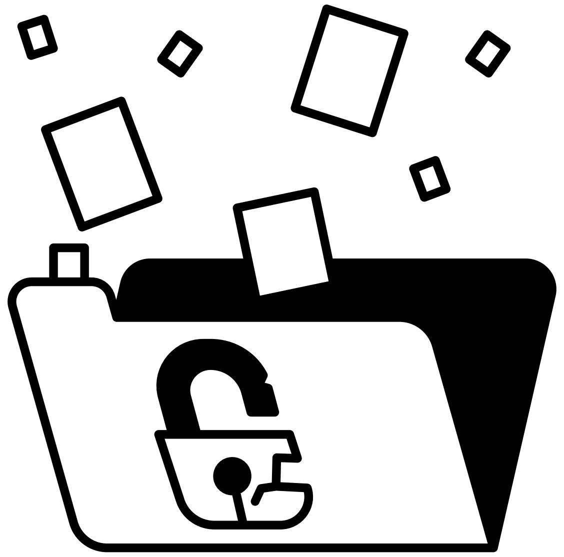
Final Output

The model is fit on the pre processed data, and **topic proportions** are obtained. Topics are allocated to each news headline, based on the topic proportion.

Example:

Sentence: Mechel increases exports from the Elga coalfield.
Topic Proportions:
International Relations & Trade: 0.523387
Saudi Arabia & Investment: 0.285559
.....
Assigned Topic: **International Relations & Trade**

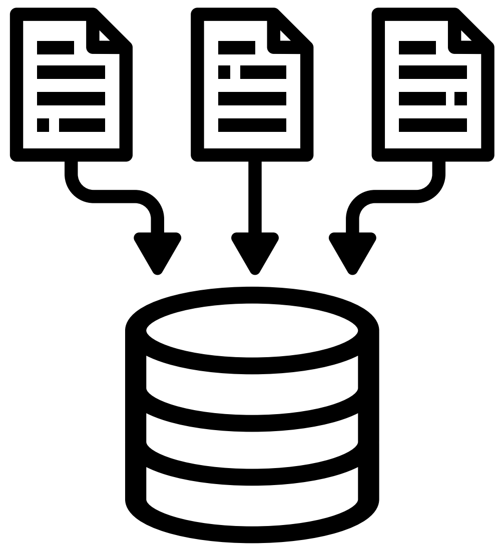
TOPIC MODELLING



Snapshot of Data:

Original Title	Energy Markets & Operations	International Relations & Trade	Environmental Sustainability	Saudi Arabia & Investment	Global Demand & China	Crude Oil Market Analysis	Highest Topic
Kenya To Make Electricity Available For 100 Percent Of Its Population	0.0484	0.2855	0.2857	0.0476	0.2851	0.0476	Topic 3 Score
An Open Letter To The U.S. Energy Secretary Nominee	0.0478	0.0476	0.0476	0.0484	0.0490	0.7596	Topic 6 Score
The South American Nation Seeing An Oil And Gold Breakout	0.1935	0.0336	0.0324	0.0323	0.5147	0.1935	Topic 5 Score
Iran Picks 29 Foreign Companies To Bid In Oil, Gas Tenders	0.0323	0.4399	0.0330	0.0329	0.4288	0.0332	Topic 2 Score
36 Killed In IS Attack In Baghdad, More Attacks On The Way	0.2857	0.5234	0.0476	0.0476	0.0481	0.0476	Topic 2 Score
Venezuela Starts 95,000 Bpd Production Cut As Part Of OPEC Deal	0.0323	0.6632	0.0324	0.0328	0.2070	0.0323	Topic 2 Score
Libya Close To 700,000 Bpd In Daily Oil Output	0.0478	0.7612	0.0480	0.0477	0.0476	0.0476	Topic 2 Score
Energy Prices Rise More Than Other Commodities In 2016	0.7584	0.0476	0.0476	0.0499	0.0476	0.0488	Topic 1 Score
Analyst: Istanbul Attack Precursor To ISIS Strike On Saudi Oil	0.0323	0.5003	0.0323	0.3706	0.0324	0.0323	Topic 2 Score
Tancoal sustains record sales	0.2856	0.0476	0.5221	0.0494	0.0476	0.0476	Topic 3 Score
Mechel increases exports from the Elga coalfield	0.0480	0.5234	0.0478	0.2856	0.0476	0.0476	Topic 2 Score
Baralaba Coal Company appoints CEO	0.0625	0.0625	0.6851	0.0634	0.0630	0.0635	Topic 3 Score
Tata Power to offset losses due to higher coal prices	0.2308	0.0385	0.2307	0.2308	0.0385	0.2308	Topic 6 Score
Goldman Sachs Sees 84% Compliance With OPEC Cuts	0.1606	0.4505	0.0278	0.3055	0.0278	0.0278	Topic 2 Score

DATA PRE PROCESSING



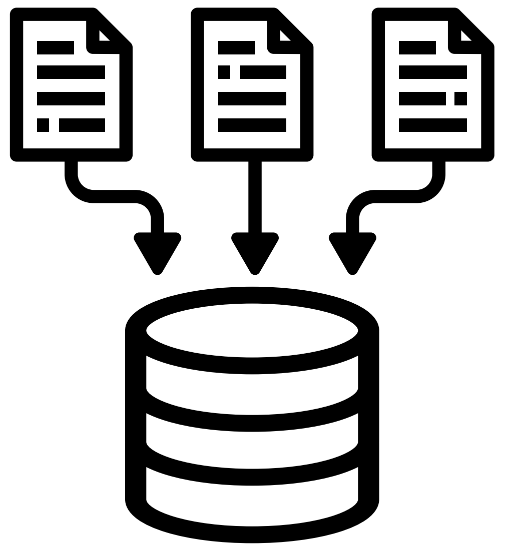
Weighted Sentiment Algorithm

Weighted average of sentiment scores is calculated to get daily sentiment scores. The Sentiment Scores of articles of a day will get more weight provided they fall in the topic which comprises of maximum articles on that day.

Concatenation

All columns are merged using the key column as **Date.** The final shape of the data is **1891** rows x **8** columns.

DATA PRE PROCESSING



MISSING VALUES

Missing Values were imputed using **Simple Moving Average.**
This Helps to smooth out fluctuations caused in the data by missing values.

SCALING

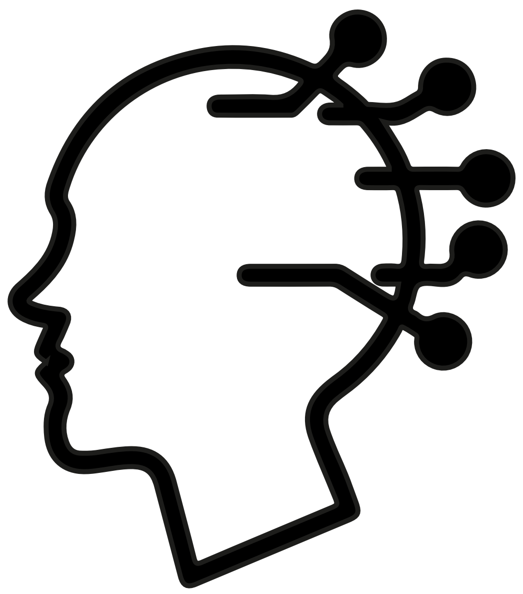
Scaled the Financial data using techniques like min-max scaling to ensure features have comparable scales.

DATA PRE PROCESSING

Snapshot of Data

Date	weighted_Polarity	weighted_Subjectivity	Richards_Bay_futures	Newcastle_futures	Dow_jones	Bloomberg_index	Argus_McCloskey_futures	Coal Price
02-01-2017	0.50	0.12	83.54	87.08	87.66	24.04	84.36	82.50
03-01-2017	0.41	0.32	84.20	90.50	86.61	23.79	82.40	84.20
04-01-2017	0.48	0.20	83.45	88.60	86.79	24.03	84.25	85.75
05-01-2017	0.50	0.44	83.75	85.70	87.04	24.22	85.65	84.75
06-01-2017	0.50	0.24	82.75	83.50	87.34	24.10	85.15	84.05
09-01-2017	0.49	0.22	82.90	82.05	86.49	23.81	84.25	86.40
10-01-2017	0.46	0.20	84.90	81.50	87.10	23.93	86.75	88.70
11-01-2017	0.46	0.29	85.65	82.15	87.12	24.09	88.55	88.60
12-01-2017	0.56	0.33	85.50	84.05	88.06	24.50	88.95	88.20
13-01-2017	0.47	0.22	85.60	83.50	88.41	24.52	88.45	88.05
16-01-2017	0.42	0.29	85.10	82.90	88.58	24.30	88.19	89.80
17-01-2017	0.47	0.31	87.75	84.55	88.97	24.63	89.75	89.95
18-01-2017	0.42	0.30	87.05	83.85	88.77	24.45	88.95	88.55
19-01-2017	0.58	0.27	86.15	83.45	88.35	24.34	88.75	89.20

NEURAL NETWORKS & LSTM

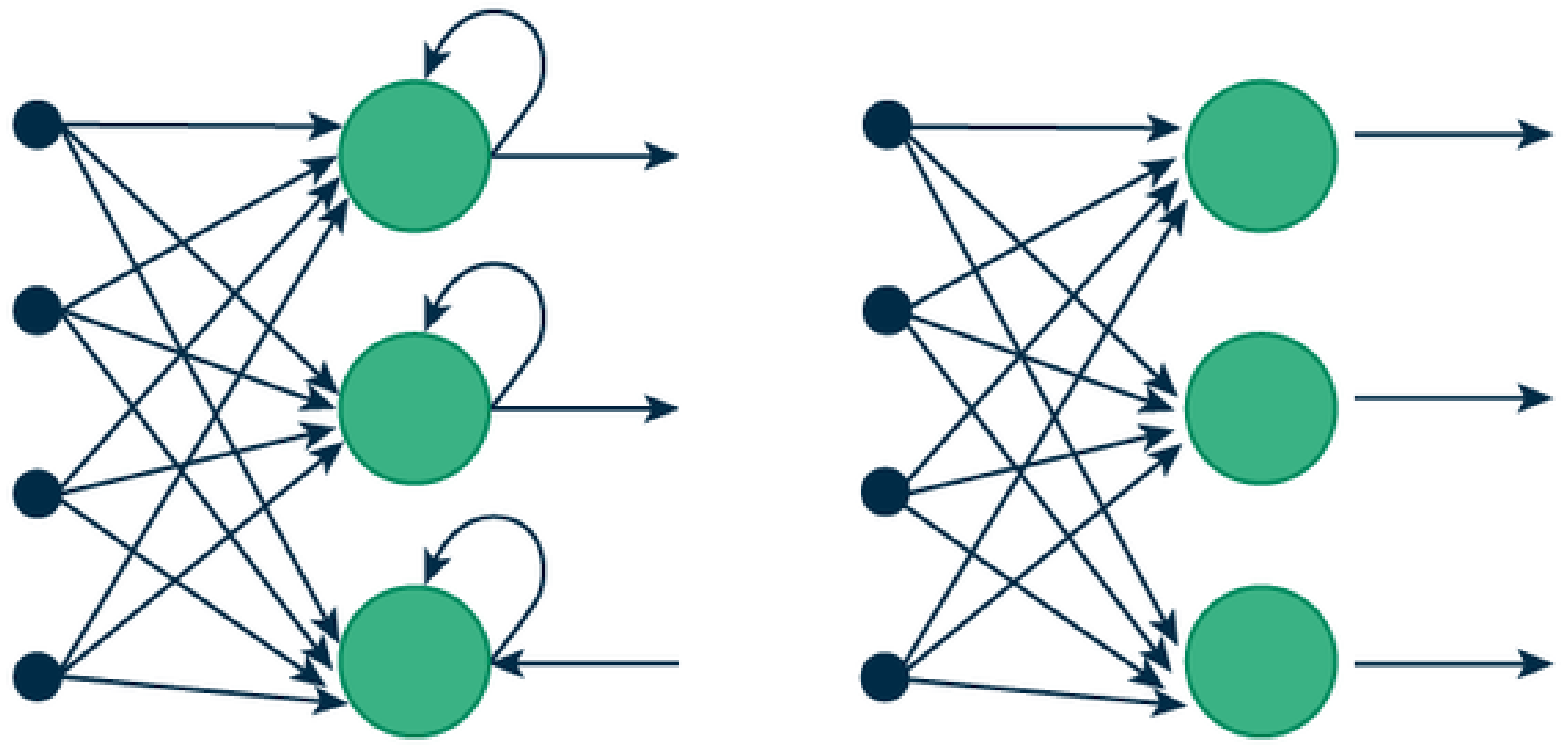


LSTM stands for **Long Short-Term Memory**, and it's a special type of neural network that's great at capturing patterns in sequences of data, like time series data or text.

LSTM is capable of learning long-term dependencies, which is its most significant advantage. It can remember information for long periods.

LSTM

Model Architecture



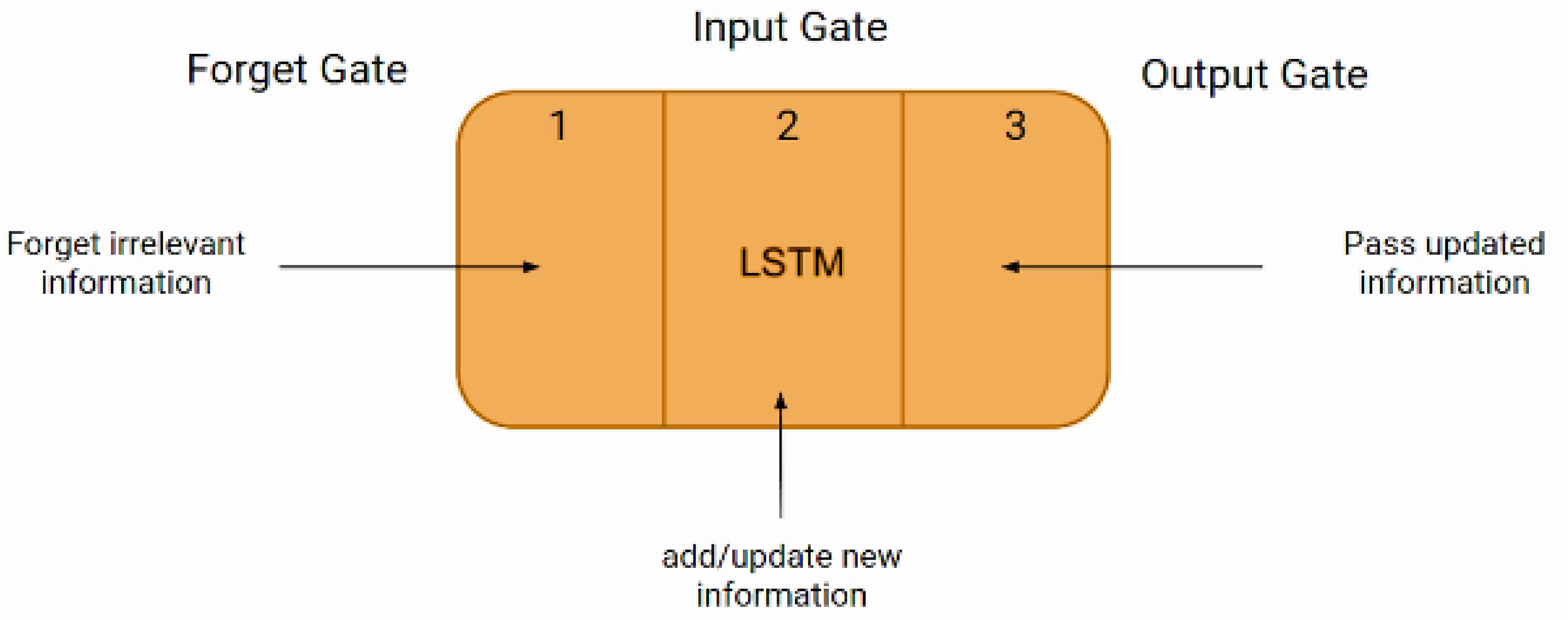
(a) Recurrent Neural Network (b) Feed-Forward Neural Network

In a **Feed-Forward Neural Network**, the input provided travels in a single direction

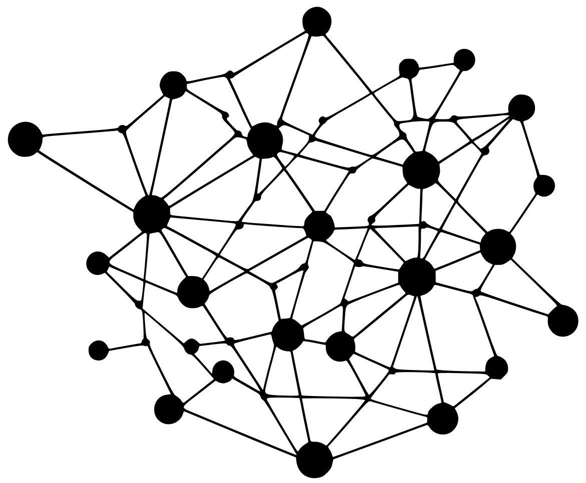
The **Recurrent Neural Network** saves the output of a layer and feeds this output back to the input to better predict the outcome of the layer

LSTM

Model Architecture



LSTM



Models:

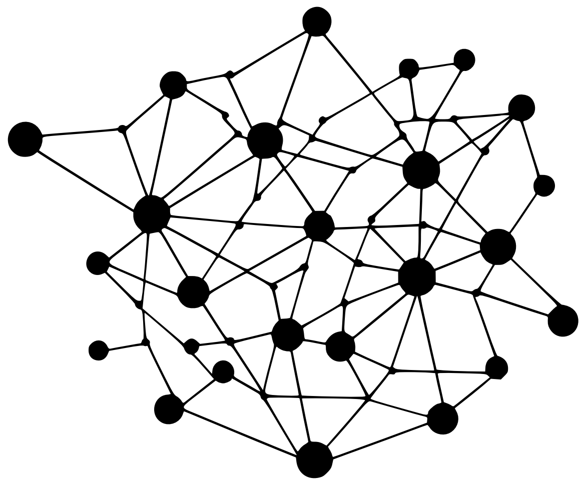
M1: Text based
This model comprises only of Sentiment Scores

M2: Financial Index
This model comprises only of Financial Indices

M3: Combined Model
All features are included in this model

Evaluation metrics such as **RMSE** and **MAE** are used to evaluate the performance of the model.

LSTM



Model Hyper Parameters

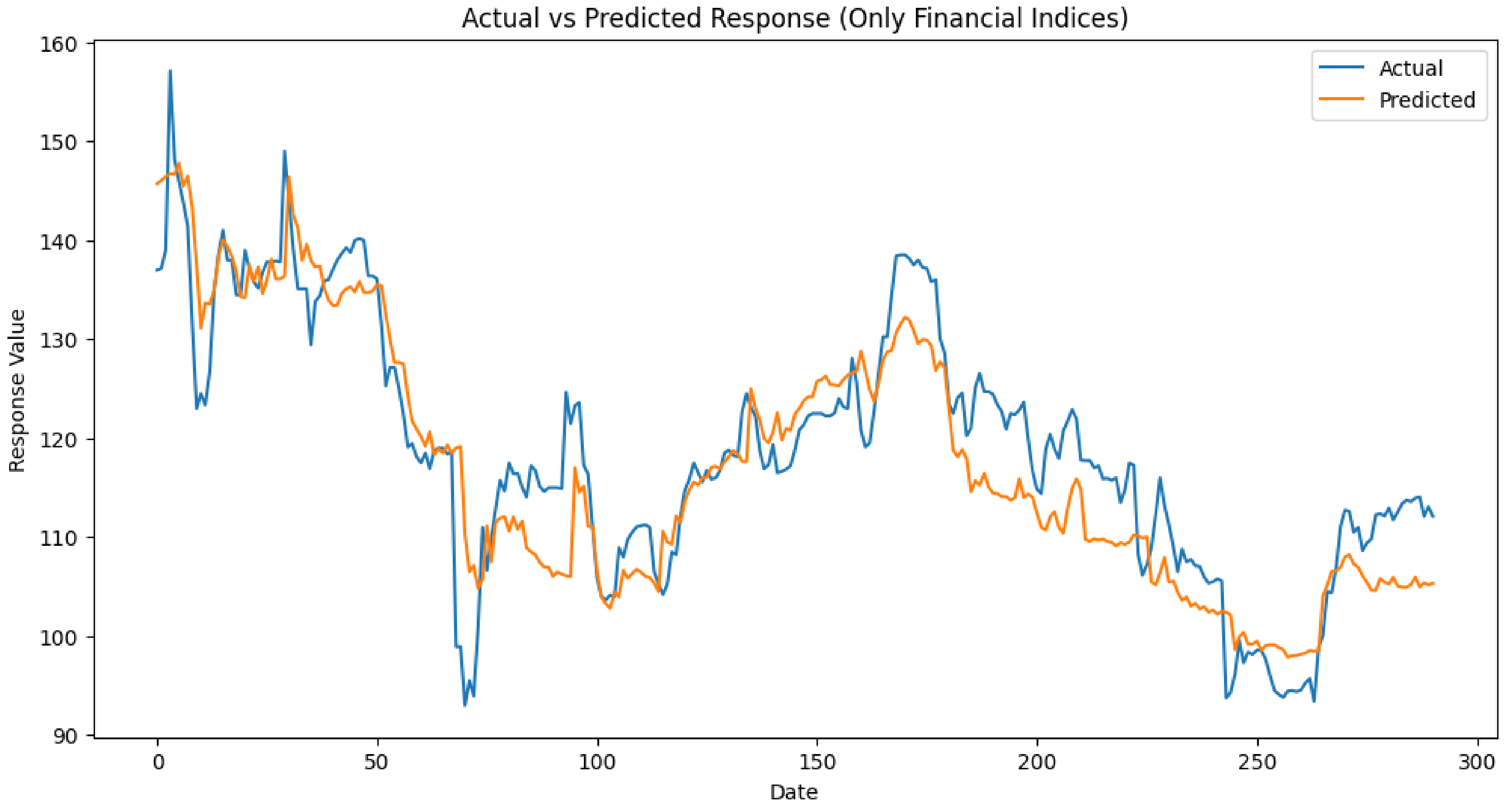
Model Description: The models were fit with 2 hidden layers and 40 neurons in each layer. Drop Out is 20%. The model is run in 32 batches for 500 epochs.

Evaluation of Models:

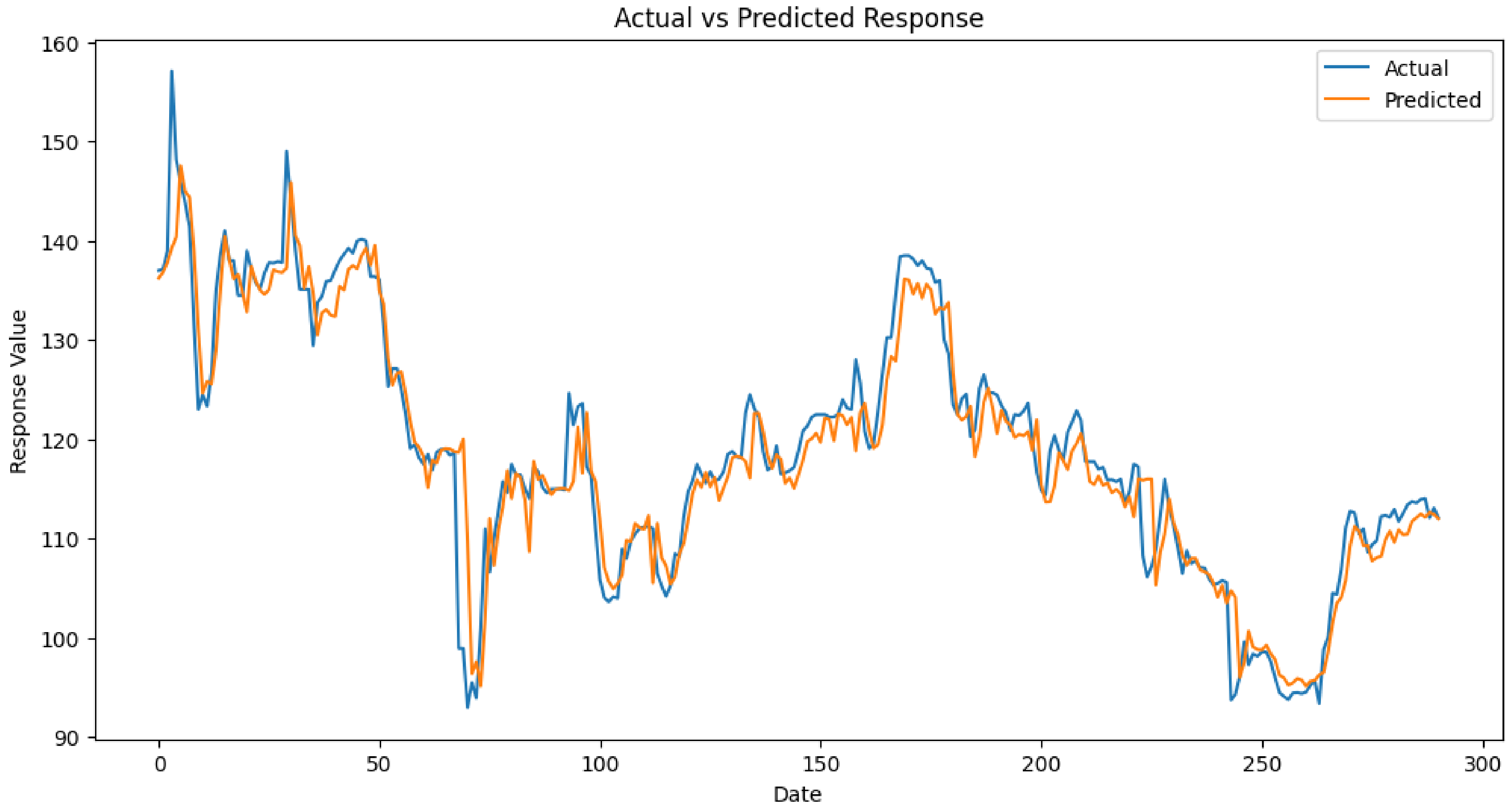
MODELS	RMSE	MAE
Text Based (M1)	16.38	13.34
Financial Indices (M2)	7.65	4.62
Combined Model (M3)	3.81	3.11

LSTM

Actual vs Predicted Graph:



Financial Index Model



Combined Model

OTHER MODELS

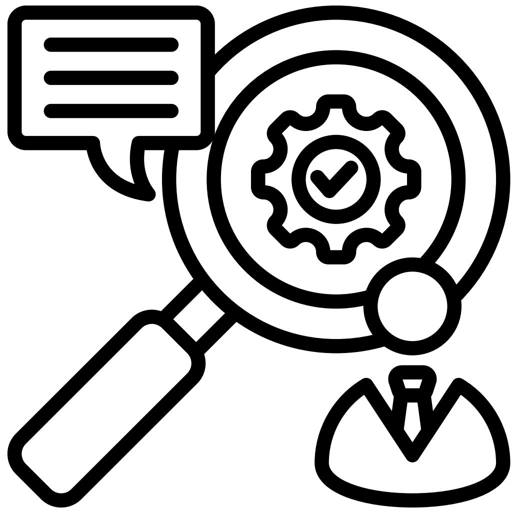
Results of LSTM are compared with two other models, viz ARIMAX and Random Forest.

RMSE	Financial Model	Combined Model	Percentage Improvement	MAE	Financial Model	Combined Model	Percentage Improvement
ARIMAX	30.145	29.972	0.57%	**ARIMAX**	26.527	26.353	0.65%
Random Forest	4.81	4.098	14.80%	**Random Forest**	3.65	3.120	14.52%
LSTM	7.65	3.81	50.19%	**LSTM**	4.62	3.11	32.68%

CONCLUSION

- **Text data** was successfully extracted for the pre-decided time period and daily average sentiment scores were obtained.
- Sentiment Scores were included in the forecasting process, and models were evaluated for the same.
- **RMSE** shows that LSTM performed the best on combined data, and the inclusion of text data reduced RMSE by 50.19%.
- **MAE** shows that LSTM performed the best on combined data, and inclusion of text data reduced MAE by 32.68%.
- The Models were able to capture the movement in the data without overfitting, and **inclusion** of Text Data showed **significant** improvement in the results.

FUTURE SCOPE



- In addition to news headlines, incorporating the summaries of entire news articles can provide a more comprehensive understanding of the market sentiments and factors influencing coal prices.
- Extending the analysis beyond traditional news sources to include authentic social media posts and trending topics might allows us to capture real-time market dynamics and emerging trends.

THANK YOU!