DATA PROPHET

TESTING...
IIT BHU

Presented by: Sanjana Garai
Aarav Mehta
Hardik Sharma

PROBLEM STATEMENT OVERVIEW

DATA PROPHET

TEAM TESTING – IIT BHU

Predict the close price of DJIA given the historical stock data and the financial news on Reddit

DATASET:

1. Open High Low Close of DJIA
2. Label and top 25 news headlines for each day

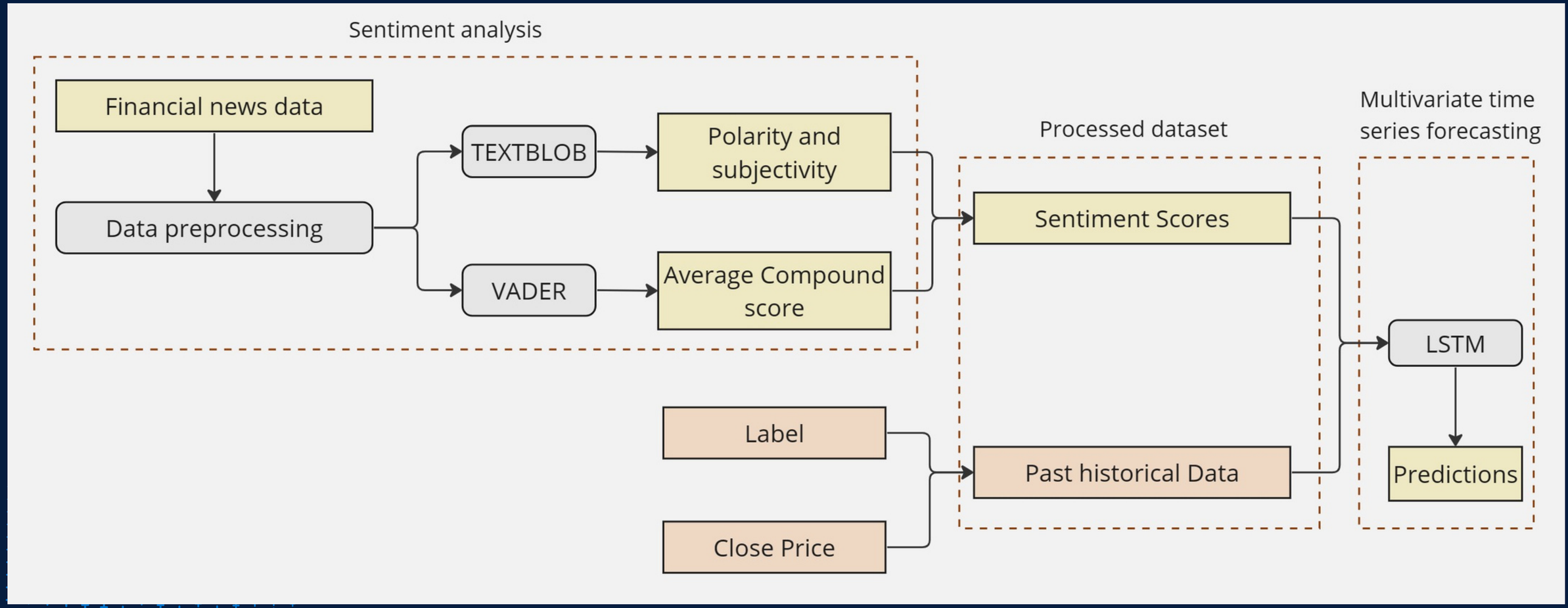
CHALLENGES:

1. Extraction of sentiments from financial news and integration with the stock price data
2. Building a model that can understand how the close price depends on different variables keeping track of past data as well

GENERAL PIPELINE

DATA PROPHET

TEAM TESTING – IIT BHU



SENTIMENT ANALYSIS

DATA PROPHET

TEAM TESTING – IIT BHU

1. Data Pre-processing: removal of NaN values and unnecessary characters before and after the string

2. Sentiment Score extraction: done using VADER and TextBlob from the NLTK library

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed for analyzing sentiments in text.

Average Compound Score

> 0 - positive emotions

< 0 – negative emotions

TextBlob

Used for processing textual data, primarily focusing on tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Polarity

degree of positivity or negativity

Subjectivity

expression of subjective opinions

VADER VS TEXTBLOB

DATA PROPHET

TEAM TESTING - IIT BHU

VADER	TEXTBLOB
Relies on a lexicon of words and phrases pre-labeled with polarity scores. It then applies valence shifting to determine the context and sentiment expressed by the entire sentence.	In addition to lexicon-based scoring, it uses other rule-based algorithms and semantic analysis like part of speech tagging or noun phrase extraction.
VADER is suitable for social media text as it accounts for the presence of punctuation marks, capital letter emphasis and slang language.	TextBlob uses the above rule-based approaches to indicate the degree to which the text expresses opinions, emotions, or personal beliefs, rather than factual information

DATASET CREATION

DATA PROPHET

TEAM TESTING – IIT BHU

1. **Prediction strategy:** Model takes feature input of 100 days and then predicts the next day's close price. We hence, divide the dataset into windows of 100 days.

2. **Feature Scaling:** The features are scaled to the range (0,1) or (-1,1) using MinMaxScaler

3. **Feature Concatenation:** All the features are then concatenated into a single dataset

Close Price | Label | Average Compound Score | Polarity | Subjectivity

4. **Fit Transform:** The dataset is resized into the format:
(num_samples, window_size, num_features)

5. **Division into training, validation and testing dataset**

And the training begins...

```
model.fit(X_train,y_train_close,validation_data=(X_val,y_val_close),epochs=100,batch_size=32,verbose=1)
```

MARKET FOREASTING

DATA PROPHET

TEAM TESTING – IIT BHU

Below are the reasons why we have chosen an LSTM for the stock market forecasting task

Required Characteristics	Features of LSTM
Sequential pattern recognition	RNN Architecture
Short term dependancies	Short term memory
Long term dependancies	Long term memory
Complex pattern recognition	Design of LSTM
Temporal sentiment analysis	RNN architecture

LSTM ARCHITECTURE

DATA PROPHET

TEAM TESTING – IIT BHU

```
model=Sequential()
model.add(LSTM(50,return_sequences=True,input_shape=(100,5)))
model.add(LSTM(50,return_sequences=True))
model.add(LSTM(50))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')
```

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_3 (LSTM)               (None, 100, 50)           11200

 lstm_4 (LSTM)               (None, 100, 50)           20200

 lstm_5 (LSTM)               (None, 50)                20200

 dense_1 (Dense)             (None, 1)                 51

=================================================================
Total params: 51651 (201.76 KB)
Trainable params: 51651 (201.76 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Hyperparameters	Value
Learning rate	0.001
Optimizer	Adam optimizer
Batch size	32
Loss function	Mean squared error
Number of epochs	100

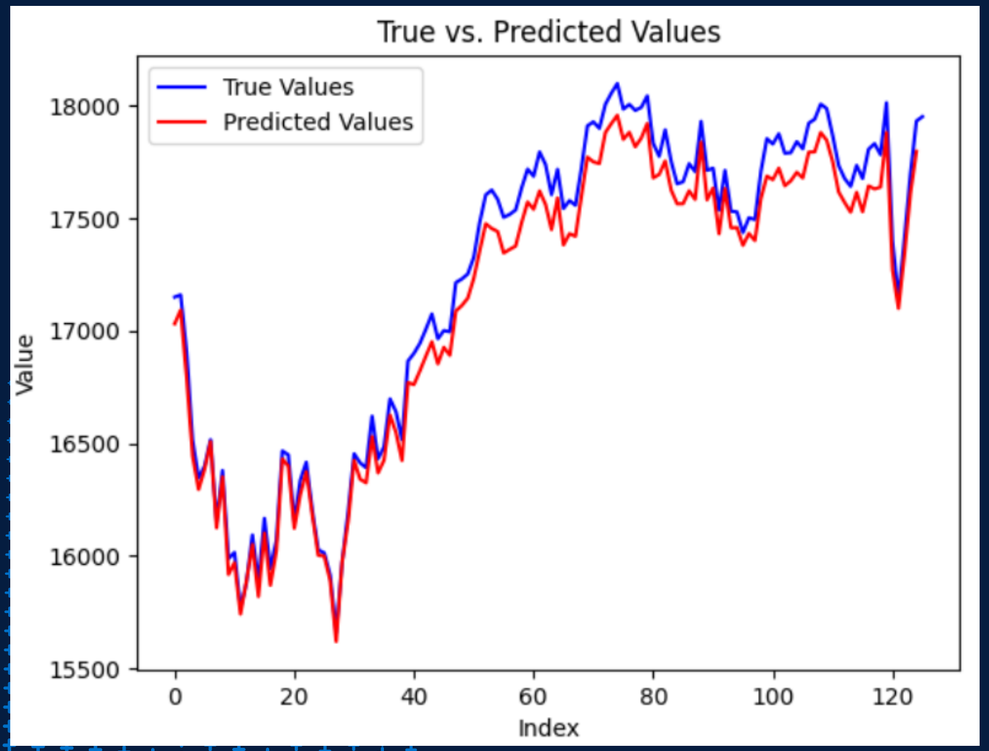
The optimal hyperparameters have been chosen using grid search

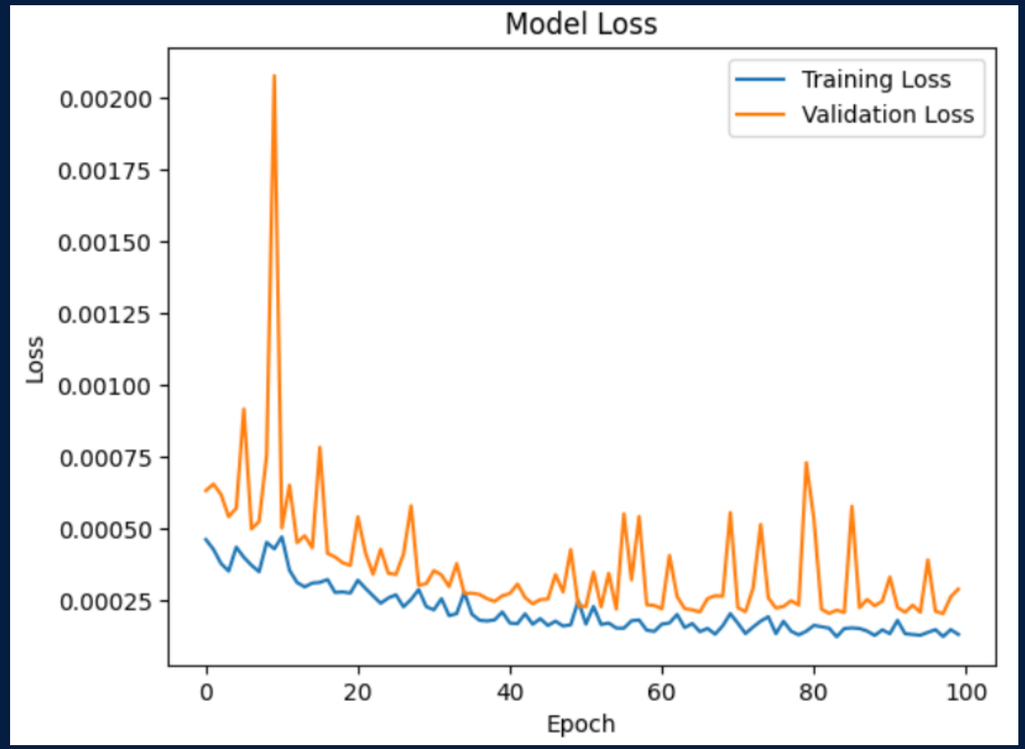
EVALUATION AND RESULTS

DATA PROPHET

TEAM TESTING – IIT BHU

Prediction loop: our dataset consists of only 1 data point which consists of the close price of the previous 100 days. We then predict the close price of the first day, append it to the dataset, append the sentiment scores and label of that day into the dataset and then remove the first entry from the dataset to predict the close price of the second day and this loop continues.





r2 score:
1 – RSS/TSS
= 0.97

*r2 score indicates how much of the variability can be explained by the model

MODEL UTILIZATION AND FURTHER IMPROVEMENT

DATA PROPHET

TEAM TESTING - IIT BHU

LSTMs are already being used in speech recognition, natural language processing and time series forecasting. However, the main drawback of LSTMs is susceptilibility to overfitting and the time taken for it to be trained.

1. **Domain - specific lexicons:** capturing finance or market specific terms

Implementation - fine-tuning a pre-trained transformer model on a dataset of financial terms

2. **Named Entity recognition -** identifying sentiments associated to specific companies, industries of financial indicators

Implementation - choose a suitable NER model and fine tune it on the dataset after suitable pre-processing

3. **Use of STEM GNN (Spectral Temporal Graph Neural Network) -** STEM GNN's works best when the data can be structure in the form of graphs. It captures both spatial and temporal features in the data

DATA PROPHET

Thank You...