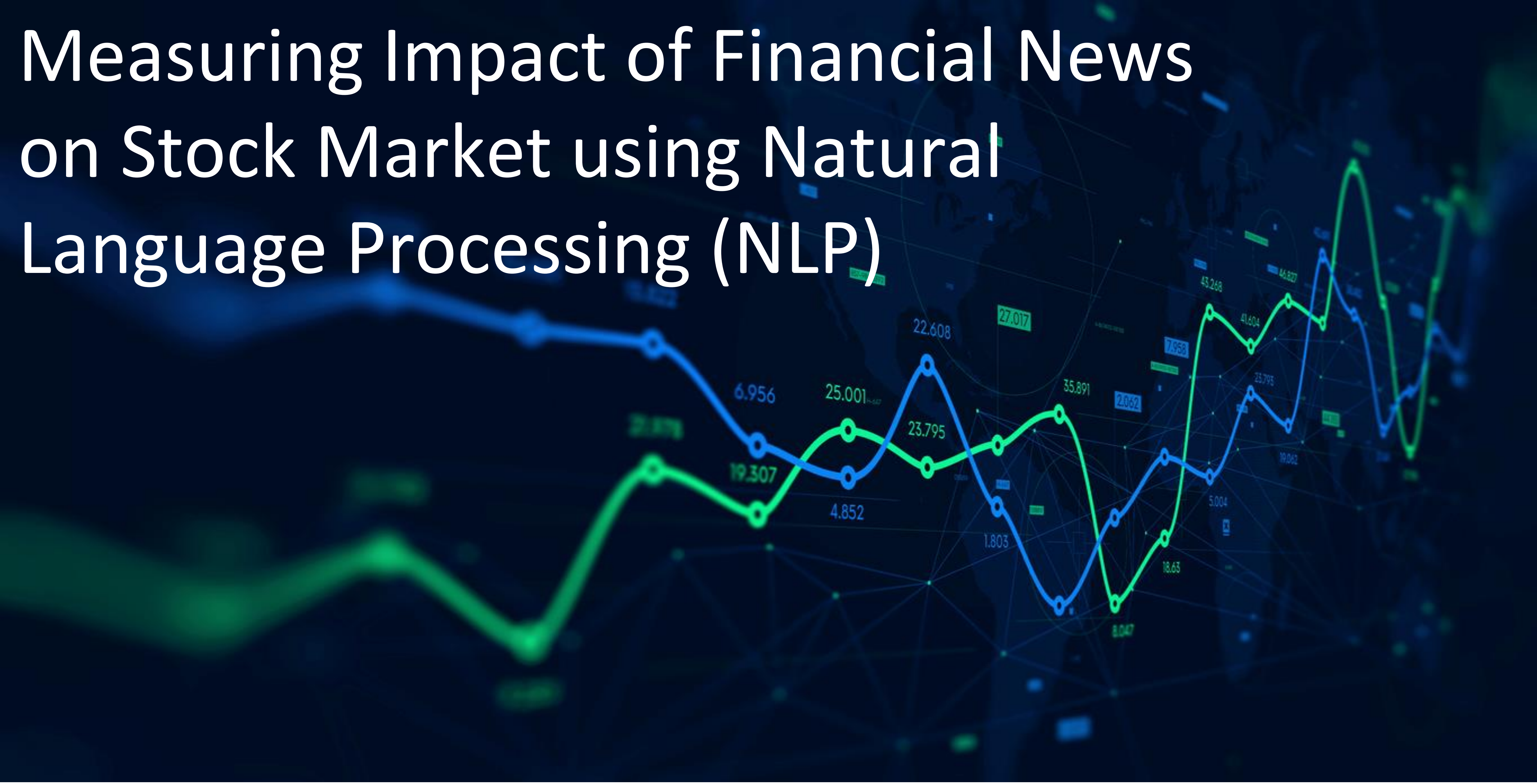
Measuring Impact of Financial News on Stock Market using Natural Language Processing (NLP)



INNODATATICS
Innovation • Data • Analytics

Team Mentors

Mr. Sharath Manikonda
Director : Innodatatics
https://www.linkedin.com/in/sharat-chandra

Mr. Sandeep Karini
Project Mentor
https://www.linkedin.com/in/sandeepkarini/

Mr. Nikhil Miryala
Data Scientist
https://www.linkedin.com/in/miryala-nikhil/

Imran Ali
Project Mentor
https://www.linkedin.com/in/mohammad-imran-ali-54a4b124/

INNODATATICS
Innovation • Data • Analytics

Team Members

Varun Sawhney

https://www.linkedin.com/in/varun-sawhney-26b7ab99

Nitin Agarwal

https://www.linkedin.com/in/nithin-agarwal-4593751a8

Harish Patil

https://www.linkedin.com/in/harish-patil-4a2173190

Bhargav Malkari

https://www.linkedin.com/in/mbhargav24

Aradyala Venkata Siva Naga Raja

https://www.linkedin.com/in/avsn-raja-770ba8214

Malika Hafiza Pasha

https://www.linkedin.com/in/malika-hafiza-pasha-1b37021a7/

Umera Pasha

https://www.linkedin.com/in/umera-pasha-25b967120

Sanjana Satpute

https://www.linkedin.com/in/sanjana-satpute-806374169

Jaya Pandey

https://www.linkedin.com/mwlite/in/jaya-pandey-7a6a61145

INNODATATICS
Innovation • Data • Analytics

Project Goals:

Objectives:

- Maximize possibility of right investment
- To measure the impact of news articles on price of the stock
- To identify the sentiment early which will help in yielding significant profit

Constraints

- Scraping the relevant information from every website is not feasible
- Minimize mixed statements and redundant data to achieve high accuracy
- Minimize misspelling, shortcuts and information duplication in the extracted textual data

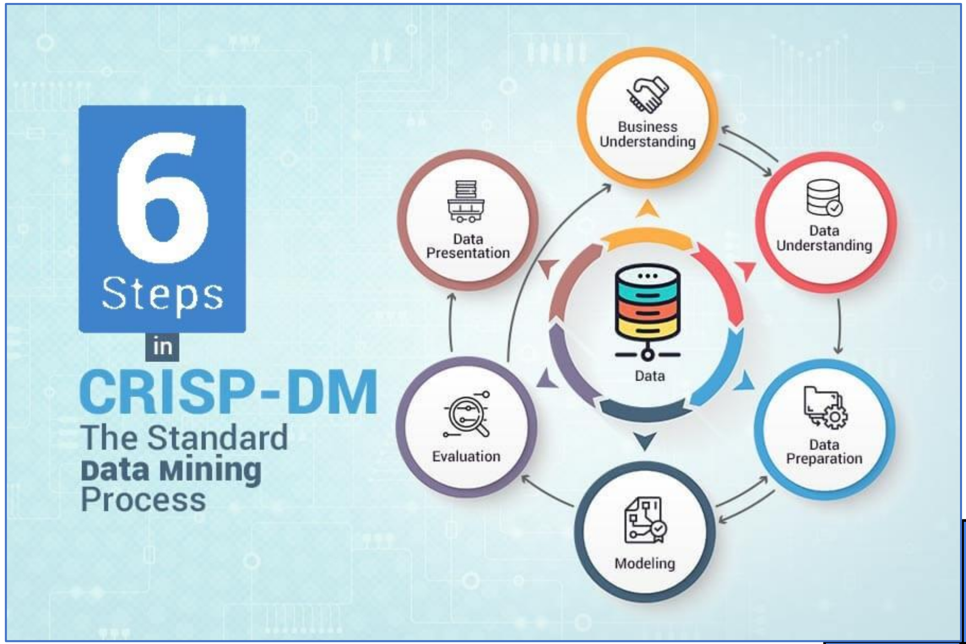
Project Overview

- The project focuses on finding sentiments of financial news and demonstrating correlation between the sentiments of the financial news and stock price variation.
- Modern behavioural finance recognizes both sentimental investors and rational investors. Manually reading the news and labelling it as positive or negative is a very tedious task. Moreover, manual evaluation of news may not be completely objective due to the factors like reader bias, emotion and fatigue. Automatic sentiment analysis can avoid these pitfalls.

Project Scope

- The main objective of the project is to refer to the overall consensus about a stock or the stock market as a whole using data from various primary sources such as Money Control, Economic Times, Yahoo Finance, Finviz, Twitter, BSE, NSE by using the Sensex and Nifty to find out whether the overall economic growth is increasing or decreasing.
- The Sentiment Analysis is advised to be done keeping the Market Measure Indicators in mind. This analysis helps to find whether the reviews are positive, negative or neutral about a specific company based on stock trading on a regular basis.

CRISP-DM Methodology



06 Deployment

05 Evaluation

04 Modelling

03 Data Preparation

02 Data Understanding

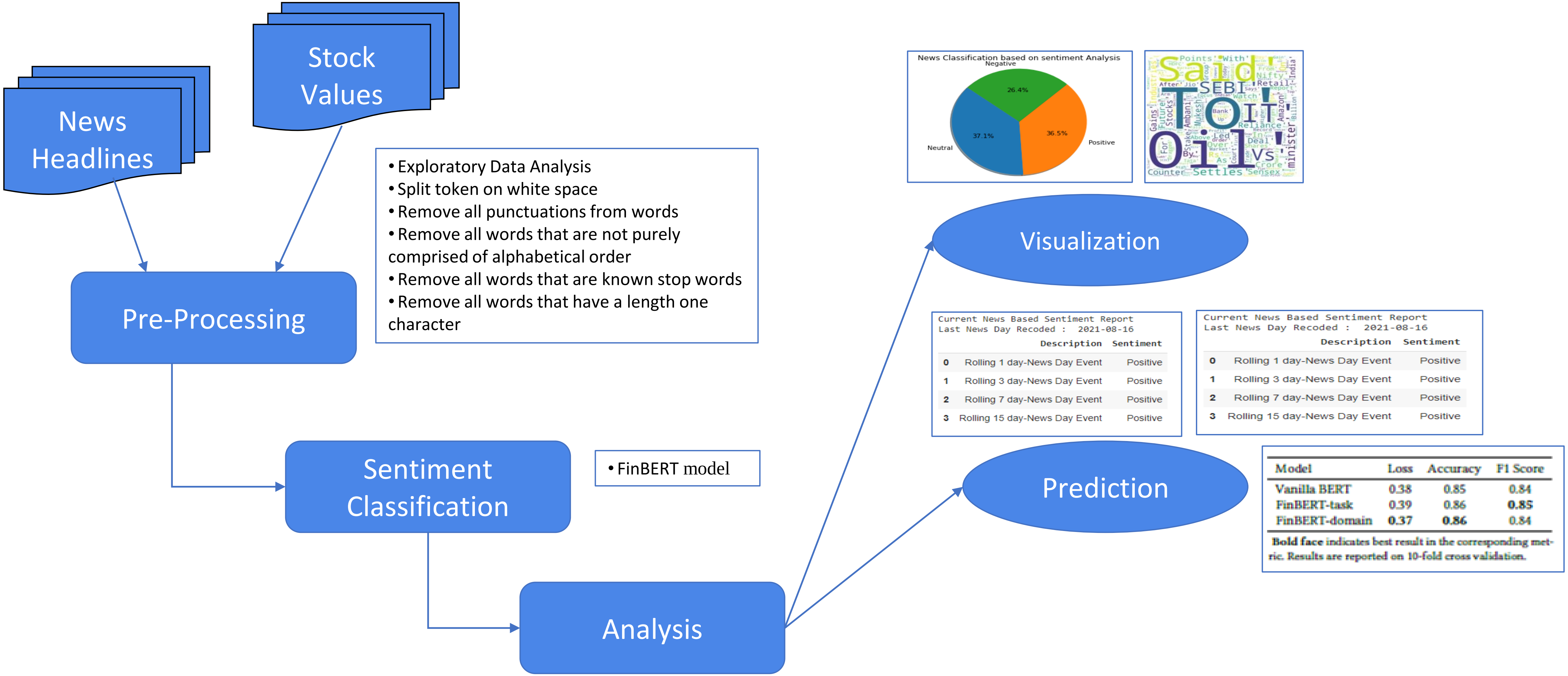
01 Business Understanding

INNODATATICS
Innovation • Data • Analytics

Success Criteria

- To provide insights to a user for making sound financial decisions.
- Easy to interpret results helps a user to understand better.
- Gains relevant information from various sources like news websites and stock prices at one platform.
- Fetches the results like news headlines in a structured format for pre-processing.
- Easy to find the shares of the company and the direction of flow using the market measure indicators with the click of a button.

Project Architecture



- Exploratory Data Analysis
- Split token on white space
- Remove all punctuations from words
- Remove all words that are not purely comprised of alphabetical order
- Remove all words that are known stop words
- Remove all words that have a length one character

- FinBERT model

```
Current News Based Sentiment Report
Last News Day Recoded :   2021-08-16
                   Description  Sentiment
0    Rolling 1 day-News Day Event   Positive
1    Rolling 3 day-News Day Event   Positive
2    Rolling 7 day-News Day Event   Positive
3   Rolling 15 day-News Day Event   Positive
```

```
Current News Based Sentiment Report
Last News Day Recoded :   2021-08-16
                   Description  Sentiment
0    Rolling 1 day-News Day Event   Positive
1    Rolling 3 day-News Day Event   Positive
2    Rolling 7 day-News Day Event   Positive
3   Rolling 15 day-News Day Event   Positive
```

Model	Loss	Accuracy	F1 Score
Vanilla BERT	0.38	0.85	0.84
FinBERT-task	0.39	0.86	**0.85**
FinBERT-domain	**0.37**	**0.86**	0.84

Bold face indicates best result in the corresponding metric. Results are reported on 10-fold cross validation.

INNODATATICS
Innovation • Data • Analytics

Technical Stacks

Language

python™

IDE

SPYDER

jupyter

CO

Libraries

NumPy

pandas

Python OS Module

Transformers

NEWS

NLTK

Se Selenium

Python import Time

RegEx Regular Expression

Strings

scikit learn

matplotlib

many expression function code operator language program Python use version example statement type name object used module

API

YAHOO! FINANCE

Web Framework

Streamlit

INNODATATICS
Innovation • Data • Analytics

System Requirement

Generic Requirements

- Textual data
- Install CUDA in the system
- For the fastest performance run it on GPU

System Requirements

- Memory: 8GB
- Graphics Card: NVIDIA GeForce GTX 970 / AMD Radeon RX 480
- CPU: Intel Core i5
- OS: Windows 8.1

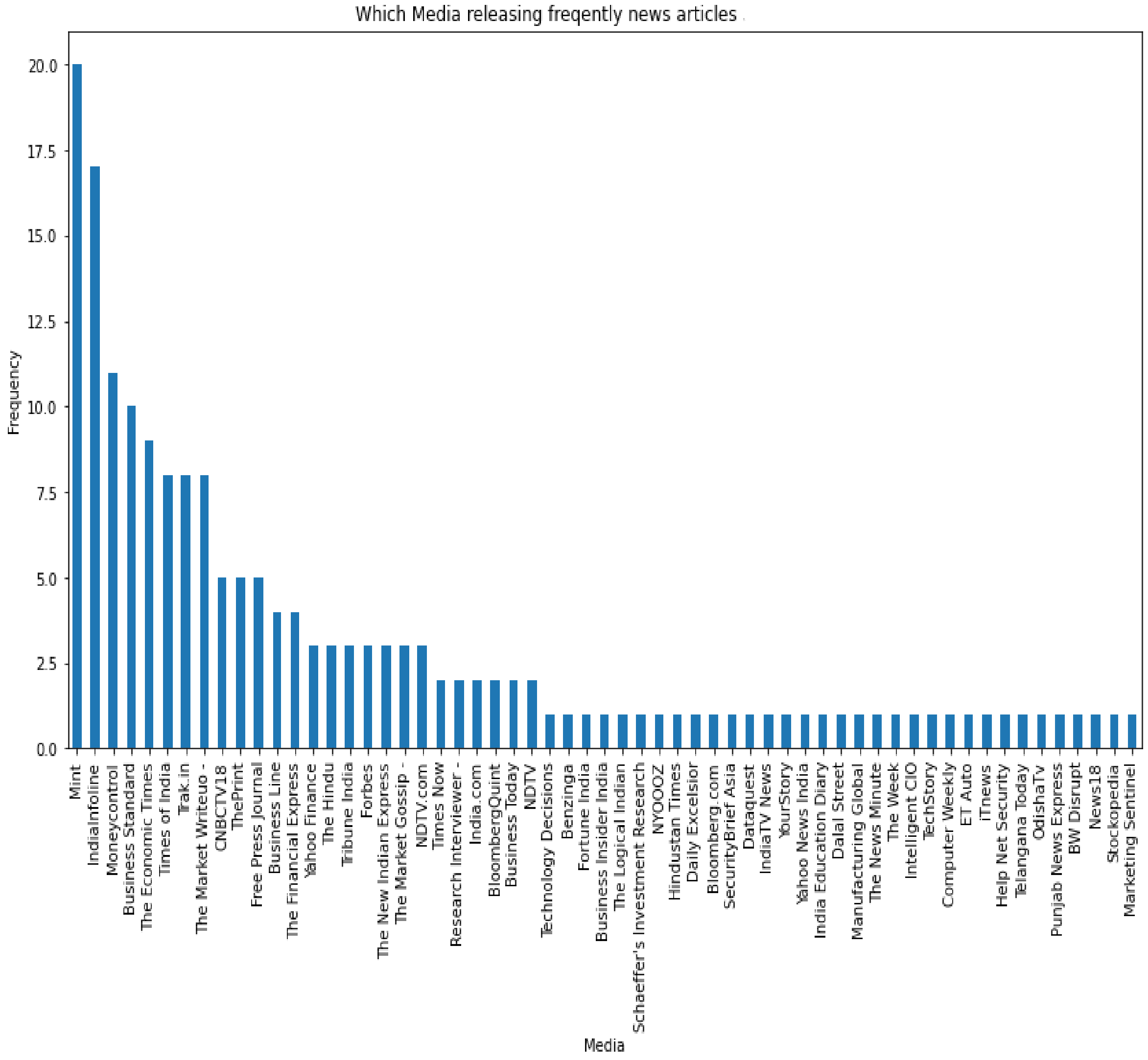
INNODATATICS
Innovation • Data • Analytics

Prepare Environment

- Have python version 3.9
- Have to install Anaconda software to launch IDE platforms like Spyder and Jupyter Notebook
- Install NLP packages like NLTK and TextBlob
- Install webdriver for Selenium package
- Install Streamlit
- Install Transformers

Data Understanding

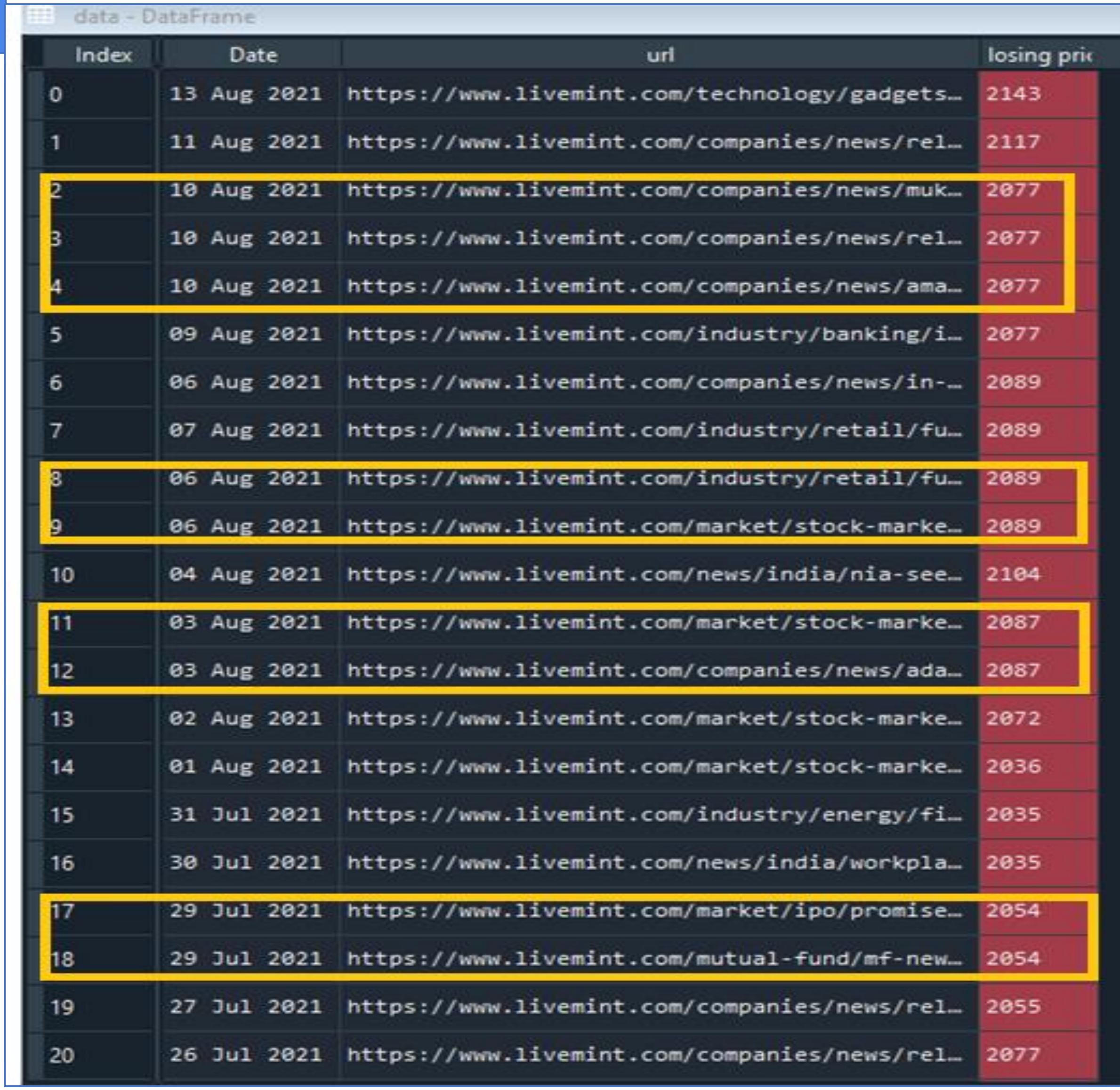
Source of the Data	News data is extracted through livemint.com and closing price data extracted through Yahoo Finance API
Type	Classification
About the Data	The data for the certain period is collected from nationalized news website and Yahoo finance website. The objective of the project is to measure the impact of news on stock price of a particular company. The news articles which are published in relation to a particular company is extracted and mapped with the values of closing price of the stock on that particular day.



INNODATATICS
Innovation • Data • Analytics

Data Collection

- Data is extracted from livemint.com and Yahoo Finance
- Problem Encountered:
  - Several news articles for a company are published simultaneously on some days.
- Steps taken to address the issue:
  - Calculated polarity for the news articles on a particular day.
  - Calculated exponential weighted average of polarity if there are several news articles for a particular day.



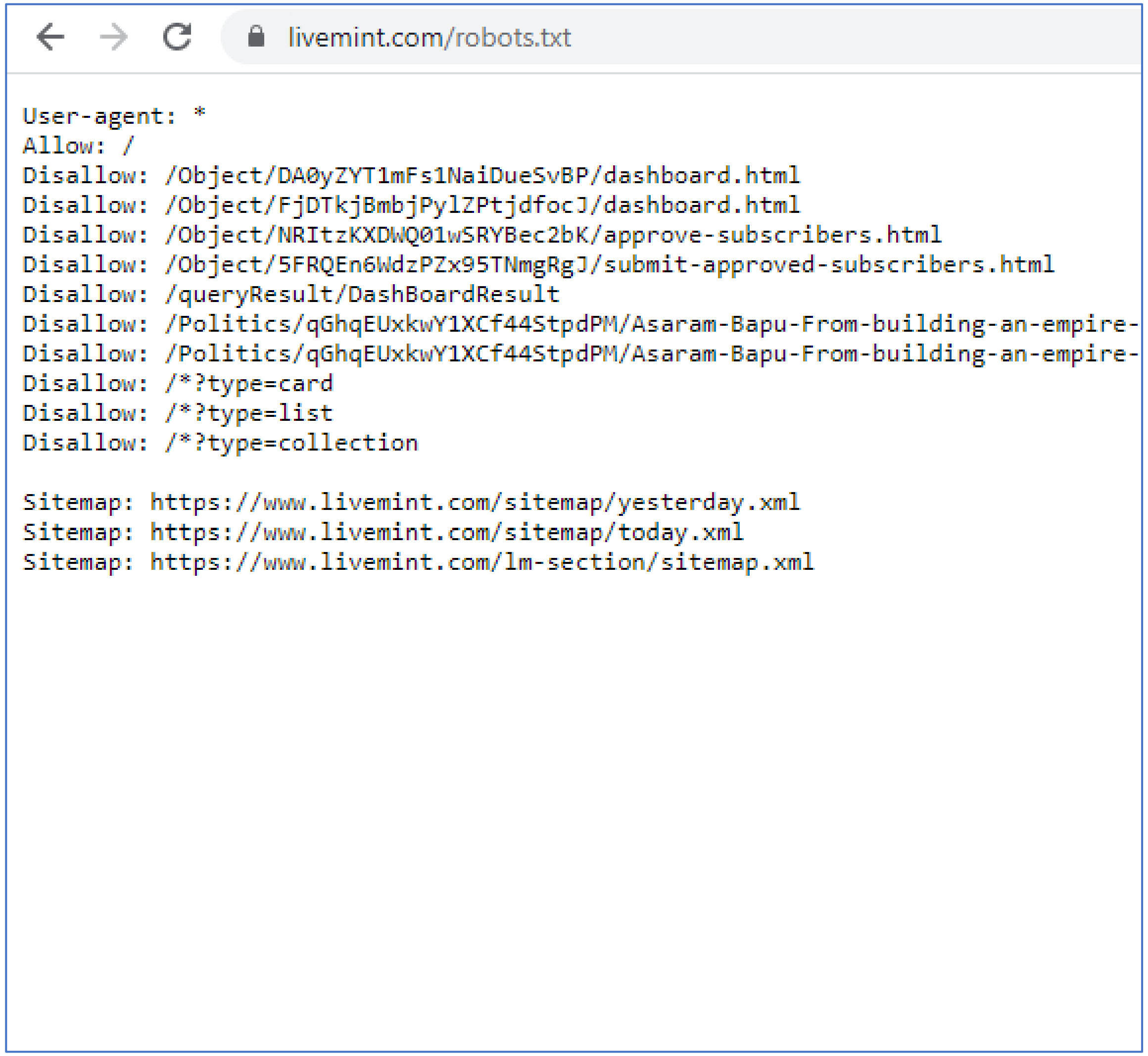
data - DataFrame

Index	Date	url	losing pric
0	13 Aug 2021	https://www.livemint.com/technology/gadgets…	2143
1	11 Aug 2021	https://www.livemint.com/companies/news/rel…	2117
2	10 Aug 2021	https://www.livemint.com/companies/news/muk…	2077
3	10 Aug 2021	https://www.livemint.com/companies/news/rel…	2077
4	10 Aug 2021	https://www.livemint.com/companies/news/ama…	2077
5	09 Aug 2021	https://www.livemint.com/industry/banking/i…	2077
6	06 Aug 2021	https://www.livemint.com/companies/news/in-…	2089
7	07 Aug 2021	https://www.livemint.com/industry/retail/fu…	2089
8	06 Aug 2021	https://www.livemint.com/industry/retail/fu…	2089
9	06 Aug 2021	https://www.livemint.com/market/stock-marke…	2089
10	04 Aug 2021	https://www.livemint.com/news/india/nia-see…	2104
11	03 Aug 2021	https://www.livemint.com/market/stock-marke…	2087
12	03 Aug 2021	https://www.livemint.com/companies/news/ada…	2087
13	02 Aug 2021	https://www.livemint.com/market/stock-marke…	2072
14	01 Aug 2021	https://www.livemint.com/market/stock-marke…	2036
15	31 Jul 2021	https://www.livemint.com/industry/energy/fi…	2035
16	30 Jul 2021	https://www.livemint.com/news/india/workpla…	2035
17	29 Jul 2021	https://www.livemint.com/market/ipo/promise…	2054
18	29 Jul 2021	https://www.livemint.com/mutual-fund/mf-new…	2054
19	27 Jul 2021	https://www.livemint.com/companies/news/rel…	2055
20	26 Jul 2021	https://www.livemint.com/companies/news/rel…	2077

Data Collection

Ethics in Web Scraping

- We have checked the livemint.com/robots.txt to know about what data is allowed by the website owner to extract and use.
- The website owner allows users to scrape news articles with following exception listed in the figure.
- This data can be used till we don’t commercialize our product.

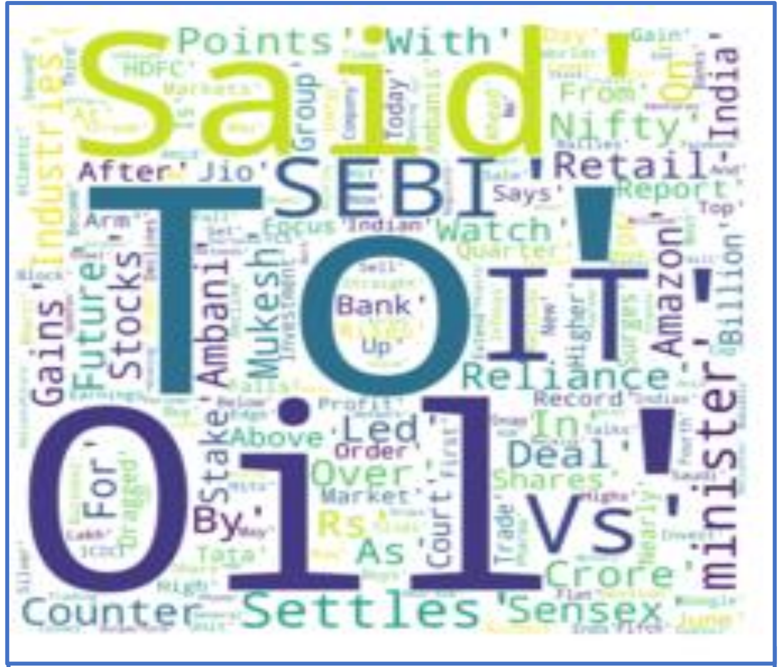


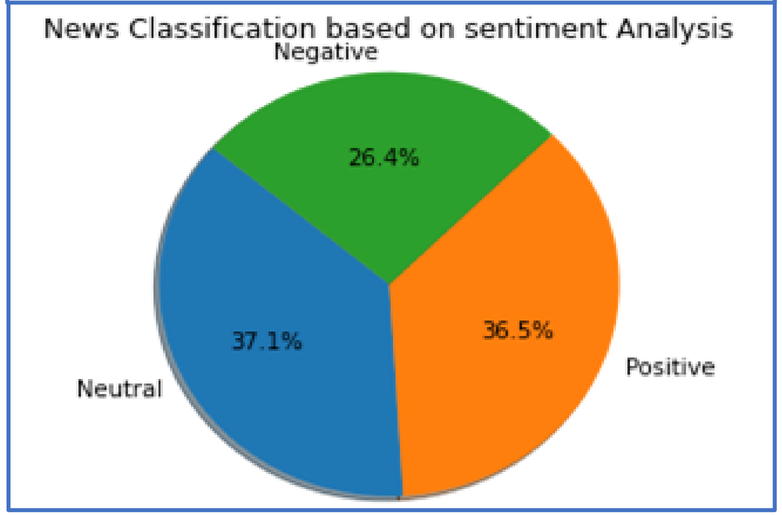
```
User-agent: *
Allow: /
Disallow: /Object/DA0yZYT1mFs1NaiDueSvBP/dashboard.html
Disallow: /Object/FjDTkjBmbjPylZPtjdfocJ/dashboard.html
Disallow: /Object/NRItzKXDWQ01wSRYBec2bK/approve-subscribers.html
Disallow: /Object/5FRQEn6WdzPZx95TNmgRgJ/submit-approved-subscribers.html
Disallow: /queryResult/DashBoardResult
Disallow: /Politics/qGhqEUxkwY1XCf44StpdPM/Asaram-Bapu-From-building-an-empire-
Disallow: /Politics/qGhqEUxkwY1XCf44StpdPM/Asaram-Bapu-From-building-an-empire-
Disallow: /*?type=card
Disallow: /*?type=list
Disallow: /*?type=collection

Sitemap: https://www.livemint.com/sitemap/yesterday.xml
Sitemap: https://www.livemint.com/sitemap/today.xml
Sitemap: https://www.livemint.com/lm-section/sitemap.xml
```

INNODATATICS
Innovation • Data • Analytics

Data Preparation

- Exploratory Data Analysis
- Data Visualization
- Split token on white space
- Remove all punctuations from words
- Remove all words that are not purely comprised of alphabetical order
- Remove all words that are known stop words
- Remove all words that have a length one character





Modelling

Unsupervised	Supervised
▪ No Labelled Data. ▪ Sentiment is predicted based on polarity of the corpus. ▪ Polarity is measured using predefined word corpus.	▪ Manually Labelling of Data related to company. ▪ News article for the stock whose price is soaring. ▪ News article for the stock whose price is dropping. ▪ Building a corpus of words that will describe the vocabulary to classify the new data.
Models : Textblob, Vader, LM Dictionary, FinBERT unsupervised model	**Models :** FinBERT supervised model

Model Selection

Model Name	Model Description	Accuracy obtained
TextBlob	The model emphasis on the interaction between the economic goods and money markets. It is based on the factors like: Investment-Demand Function, Consumption Function, Money Demand Function, Supply of Money.	**52%**
VADER	It is a model used for text sentiment analysis that is sensitive to both polarity and intensity of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. VADER relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.	**44%**
LM Dictionary	The model emphasizes the interaction between the economic goods and money markets. It is based on the factors like: Investment-Demand Function, Consumption Function, Money Demand Function, Supply of Money.	**58%**
FinBERT	The model is further extension of state of the are BERT model and is further trained on financial corpus. It is currently the state of the are model for financial data sentiment classification.	**85%**

INNODATATICS
Innovation • Data • Analytics

FinBERT Model

As we Know due to lack of labeled financial datasets, it is difficult to use any supervised model for sentiment analysis. FinBERT is a language model based on BERT specifically designed to tackle this problem.

For FinBert, BERT language model is further trained on a subset of Reuters TRC2 dataset and Financial PhraseBank to achieve this task.

BERT has two factors, which enables it to achieve the state-of-the-art results in multiple NLP tasks.:

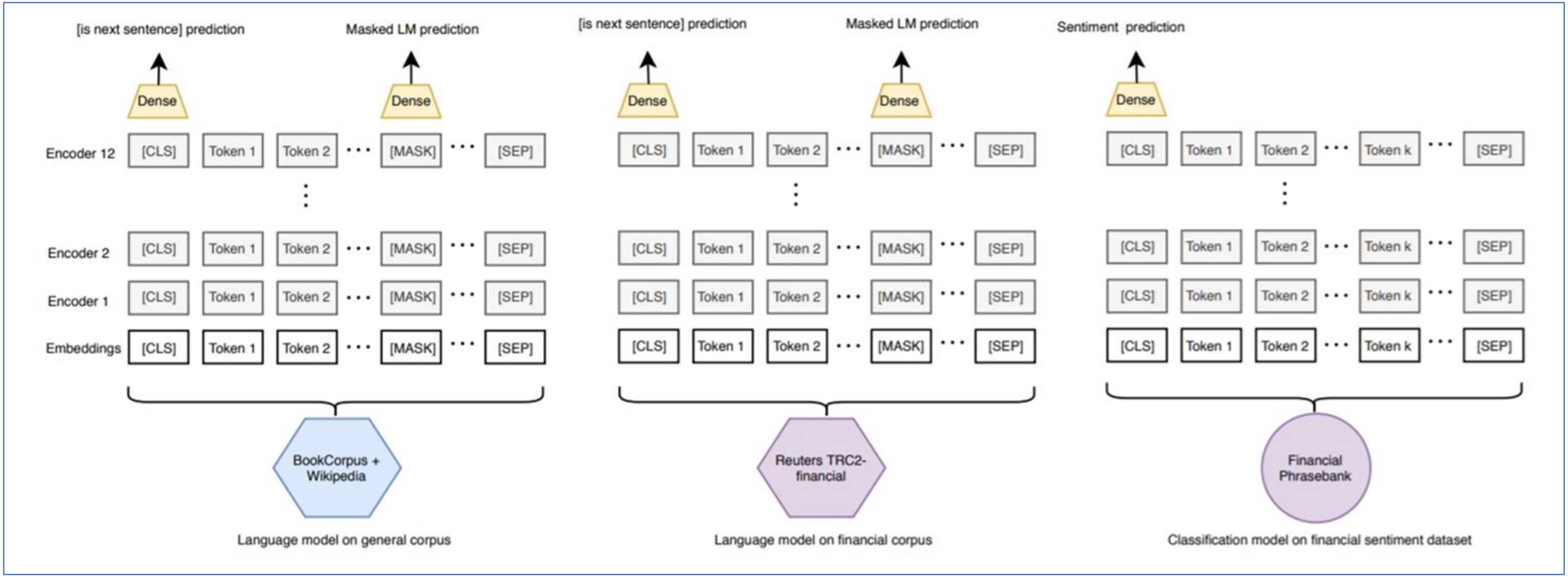
1) It defines the task of language modeling as predicting randomly masked tokens in a sequence rather than the next token, in addition to a task of classifying two sentences as following each other or not.

2) It is a very big network trained on an unprecedentedly large corpus.

In BERT, the input sequence is represented with token and position embeddings. Two tokens denoted by [CLS] and [SEP] are added to the beginning and end of the sequence respectively. For all classification tasks, including the next sentence prediction, [CLS] token is used.

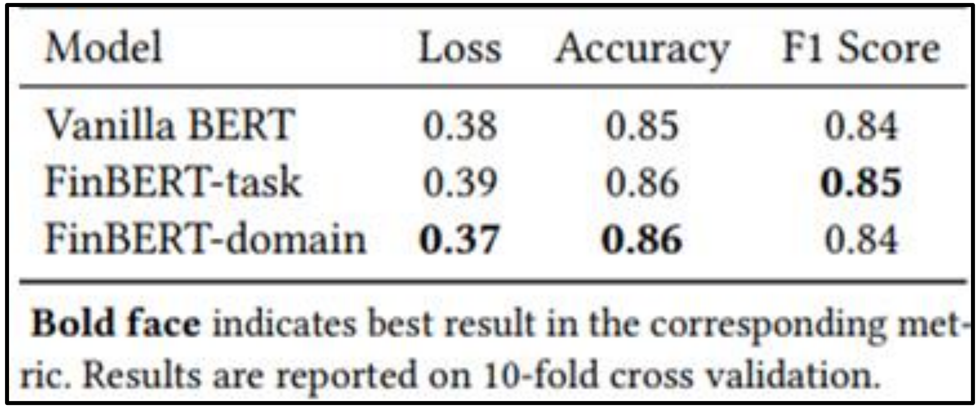
Sentiment classification in FINBert is conducted by adding a dense layer after the last hidden state of the [CLS] token. This is the recommended practice for using BERT for any classification task. Then, the classifier network is trained on the labeled sentiment dataset.

Training Overview



Model Evaluation

- We have manually labelled the data to check the accuracy of how our model is behaving with the real-time generated data.
- We have labelled data extracted from livemint.com for Reliance industries and checked model performance on 392 news headlines.
- The various models that we have used in our analysis are TextBlob, Vader sentiment analysis, LM dictionary and FinBERT.
- Out of all models, FINBERT is showing maximum accuracy.
- To Further improve accuracy, we need to train our model on large corpus of companies data that we extract from the news source.



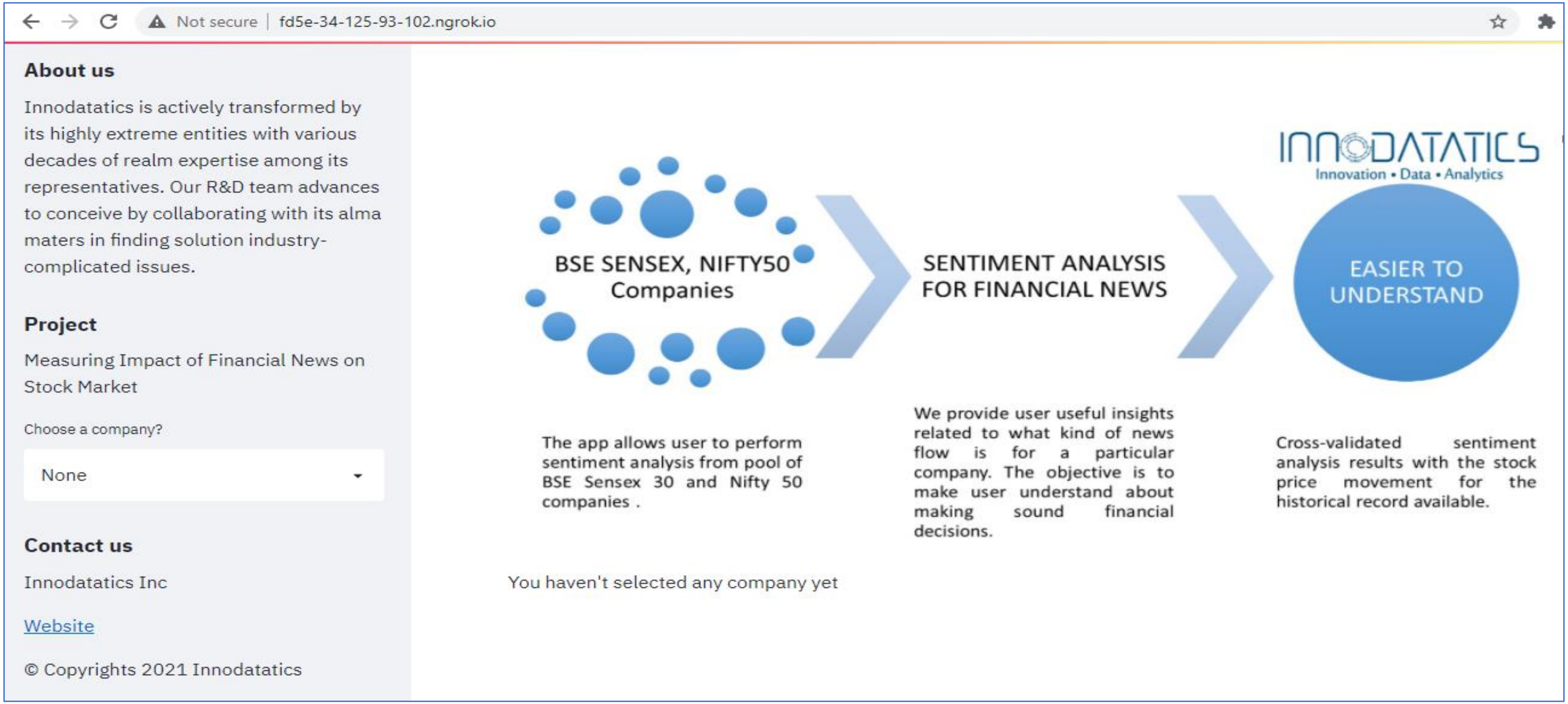
Model	Loss	Accuracy	F1 Score
Vanilla BERT	0.38	0.85	0.84
FinBERT-task	0.39	0.86	**0.85**
FinBERT-domain	**0.37**	**0.86**	0.84

Bold face indicates best result in the corresponding metric. Results are reported on 10-fold cross validation.

Requirements to Deploy Application

- Initialize a project
- Create a project skeleton
- Add the pre-trained model and create an interface to abstract the inference logic.
- Update the request handler function to return predictions using the model.
- Start the server and send a test request.

Input



About us

Innodatatics is actively transformed by its highly extreme entities with various decades of realm expertise among its representatives. Our R&D team advances to conceive by collaborating with its alma maters in finding solution industry-complicated issues.

Project

Measuring Impact of Financial News on Stock Market

Choose a company?

None

Contact us

Innodatatics Inc

Website

© Copyrights 2021 Innodatatics

BSE SENSEX, NIFTY50 Companies

SENTIMENT ANALYSIS FOR FINANCIAL NEWS

INNODATATICS Innovation • Data • Analytics

EASIER TO UNDERSTAND

The app allows user to perform sentiment analysis from pool of BSE Sensex 30 and Nifty 50 companies .

We provide user useful insights related to what kind of news flow is for a particular company. The objective is to make user understand about making sound financial decisions.

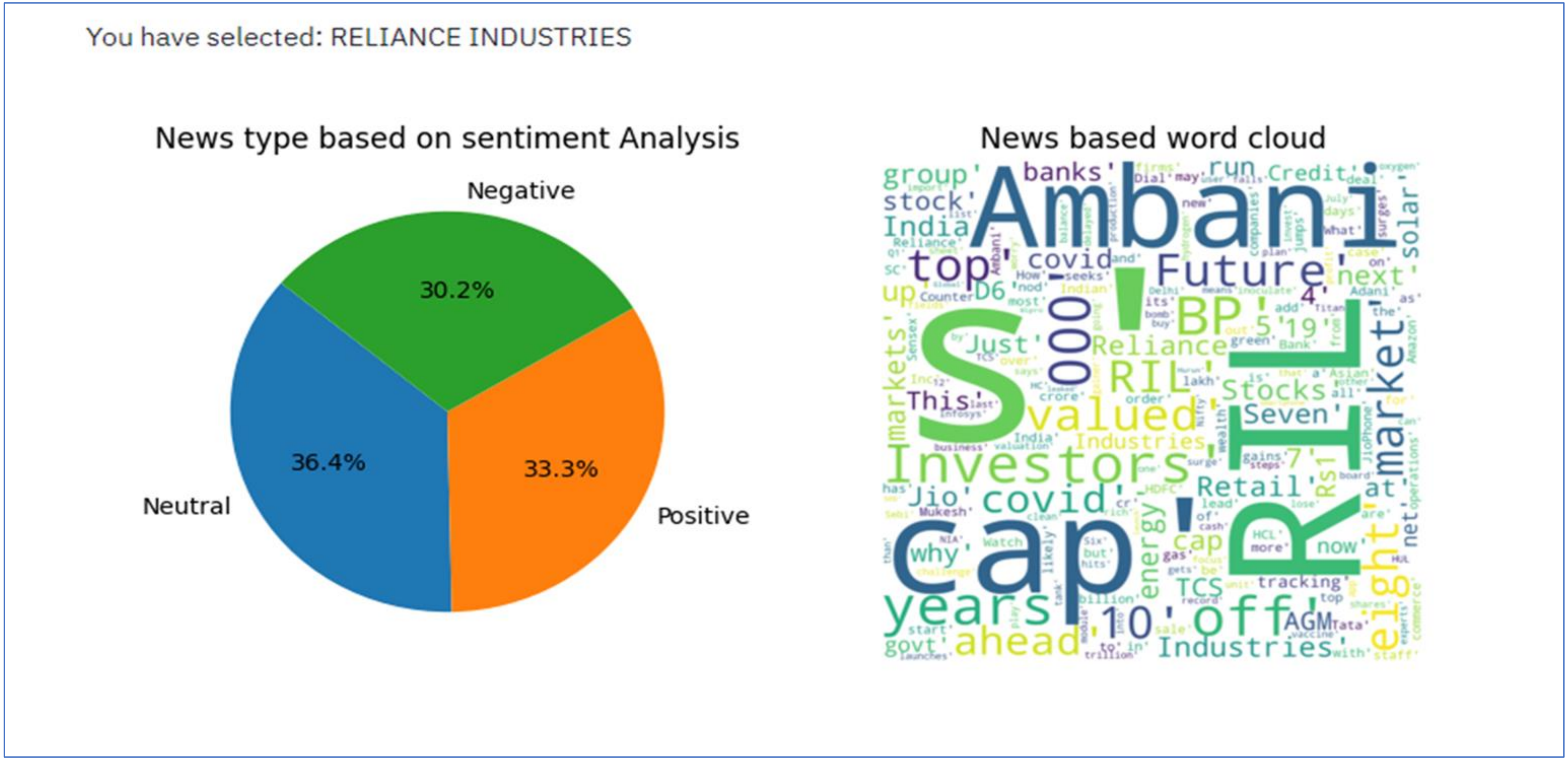
Cross-validated sentiment analysis results with the stock price movement for the historical record available.

You haven't selected any company yet

INNODATATICS Innovation • Data • Analytics

Output

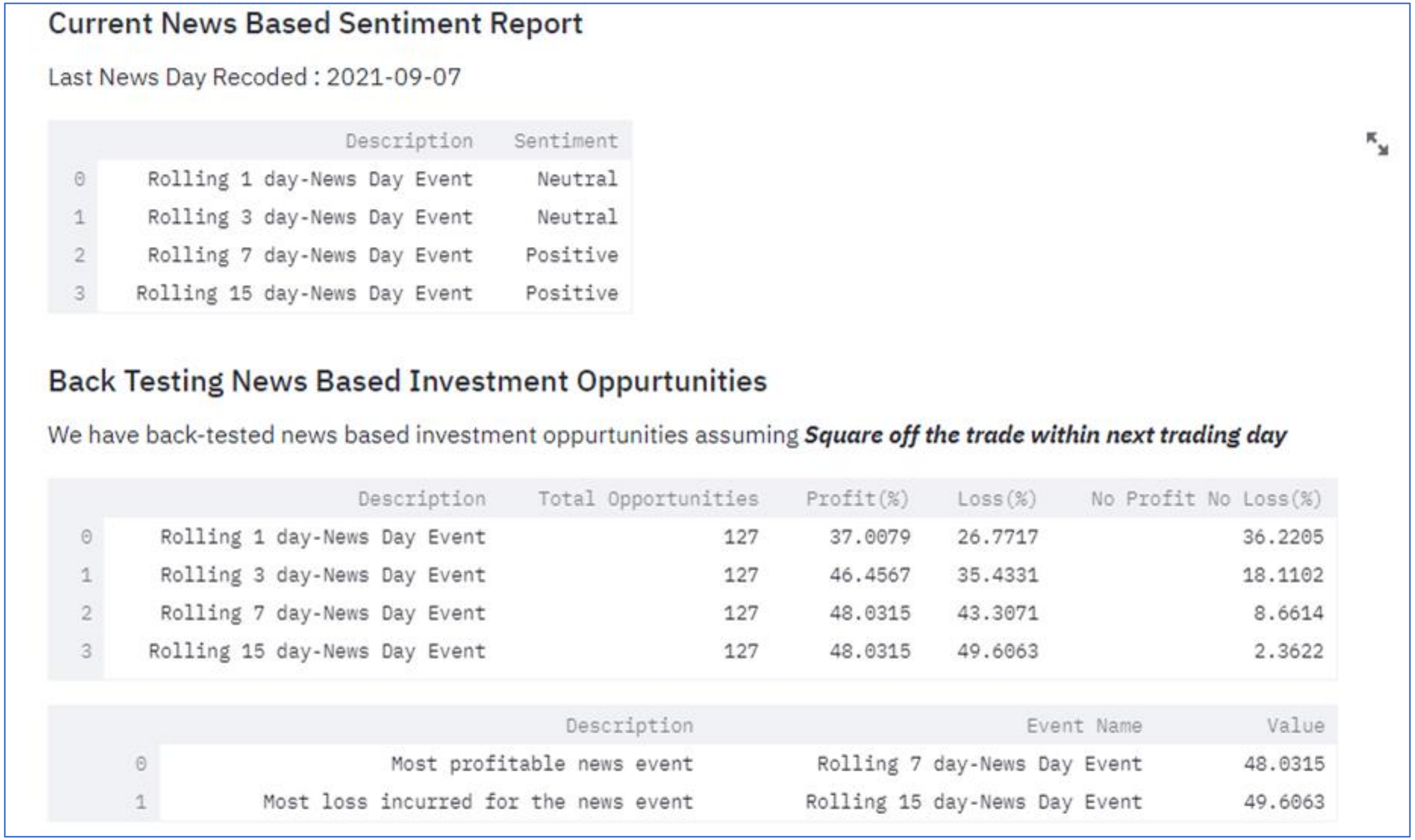
Output when a user selects a particular company:



You have selected: RELIANCE INDUSTRIES

Output

Insights generated for the User:



Current News Based Sentiment Report

Last News Day Recoded : 2021-09-07

	Description	Sentiment
0	Rolling 1 day-News Day Event	Neutral
1	Rolling 3 day-News Day Event	Neutral
2	Rolling 7 day-News Day Event	Positive
3	Rolling 15 day-News Day Event	Positive

Back Testing News Based Investment Oppurtunities

We have back-tested news based investment oppurtunities assuming ***Square off the trade within next trading day***

	Description	Total Opportunities	Profit(%)	Loss(%)	No Profit No Loss(%)
0	Rolling 1 day-News Day Event	127	37.0079	26.7717	36.2205
1	Rolling 3 day-News Day Event	127	46.4567	35.4331	18.1102
2	Rolling 7 day-News Day Event	127	48.0315	43.3071	8.6614
3	Rolling 15 day-News Day Event	127	48.0315	49.6063	2.3622

	Description	Event Name	Value
0	Most profitable news event	Rolling 7 day-News Day Event	48.0315
1	Most loss incurred for the news event	Rolling 15 day-News Day Event	49.6063

Limitations

- The foremost limitation faced in this work is the chaotic market behaviour.
- The key assumption of news-based sentiment analysis is that the news moves the markets. While news does move the markets, the markets may move without news too.
- The second limitation was the extraction of event-related information from the news stories using various models like TextBlob, Vader, LM Dictionary which did not account for the word order and other semantic roles of news text.
- For this project, only the news headlines are being considered instead of considering the content of the news article. This can sometimes result into inappropriate conclusions.
- Similar company name, may create a problem during data extraction process.

Conclusion

- Financial markets behaviour is an outcome of the prevailing sentiment shaped by the dynamics of the arriving news. Often, markets take time to decipher the sentiment. An investor who identifies the sentiment early would significantly profit from the anticipated direction.

- FinBERT model performed best with market measure indicators in the context of the Indian market. This was further validated by the manually labelled data. The results are consistent and indicate that the buying behaviour leads the spot market sentiment.

- This project captures the impact of real-time news on the markets and have significant implications for future trading, especially for high-frequency algorithmic trading. It also establishes that we can further extend our research to check the arbitrage window of opportunity in the Indian markets, which underlines the importance of high-frequency analysis of news on the markets.

Reference

- IIMB Research Paper https://www.sciencedirect.com/science/article/pii/S0970389619301569
- Federal Reserve Bank of San Francisco https://www.frbsf.org/economic-research/files/wp2017-01
- FinBERT Paper - https://arxiv.org/abs/1908.10063

INNODATATICS
Innovation • Data • Analytics