CMPT 733 - Big Data Programming 2
Project Report

R.I.S.K : Revolutionize Investment Strategies using KPIs

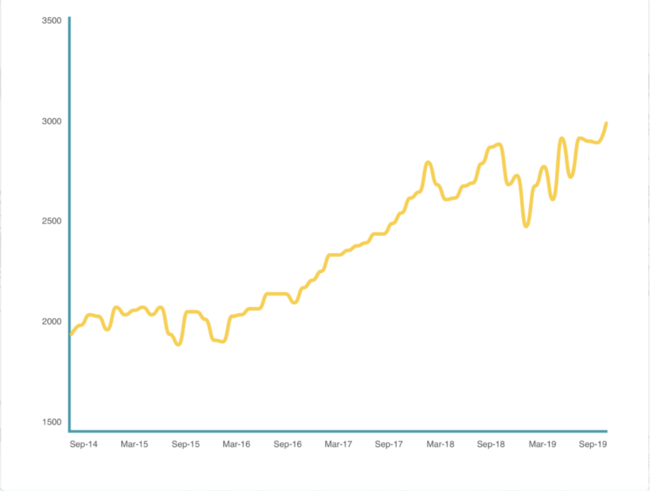
Group Name: Bloggers

Ravi Kiran Dubey, Ruchita Robert Rozario,
Slavvy Johnson Coelho, Ziyue Xia

Content

1. Motivation and Background

In the past five years, the U.S Stock Market has returned 53.1% of its investments. [1] You can use the stock market to create a steady passive income stream and build lifelong-wealth that will ensure financial security for the rest of your life, starting right now. The idea is to ‘Make your money work for you instead of you working so hard for the money’.



Imagine yourself as an investor looking to make your next big investment in the financial market. What would be the factors you’d consider to make the right choice? Today the internet is a vast data pool and extracting relevant information to answer your problem statement can be a challenging task. Data science tools and algorithms aid this process. The finance industry can leverage these technologies to make an impact on the revenues.

Our data product aims at creating a one stop platform for key investors in the industry to garner insightful information about potential companies that they can invest in and thus make a profitable decision.

2. Problem Statement

We intended on supporting the investors decision by analysing the companies' Key Performance Indicators (KPIs) over the past few months. The KPIs are evaluation metrics for the company's progress in form of statistics and strong numbers that give a clear idea about its growth trajectory. The KPIs that we selected were:

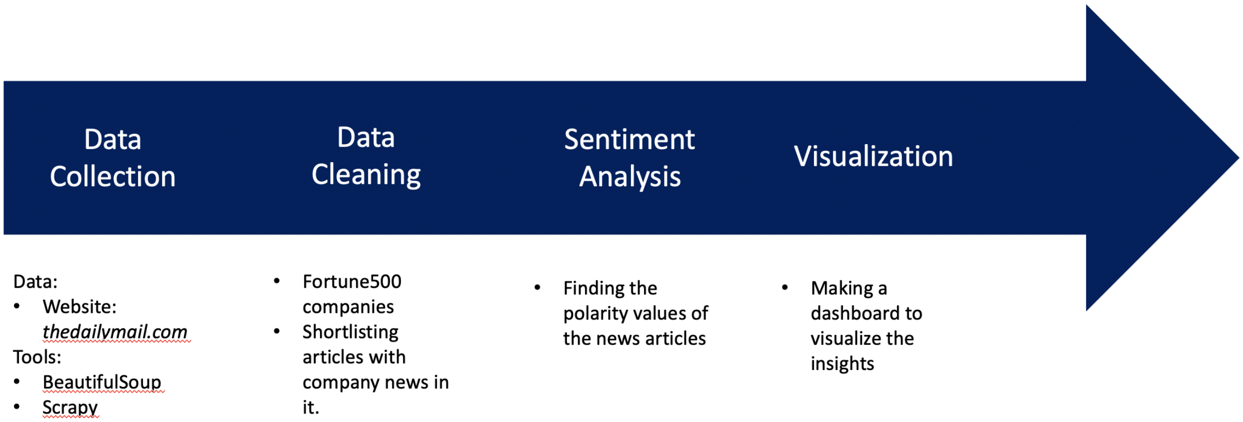
- Market Sentiment Analysis
- Stocks
- Sector wise Analysis and financial KPIs

3. Data Science Pipeline

Each KPI considered in the project has an extensive pipeline of its own. In this section, we will proceed to explain each KPI with its Data Science Pipeline. We combined them in the end to get a final Data product.

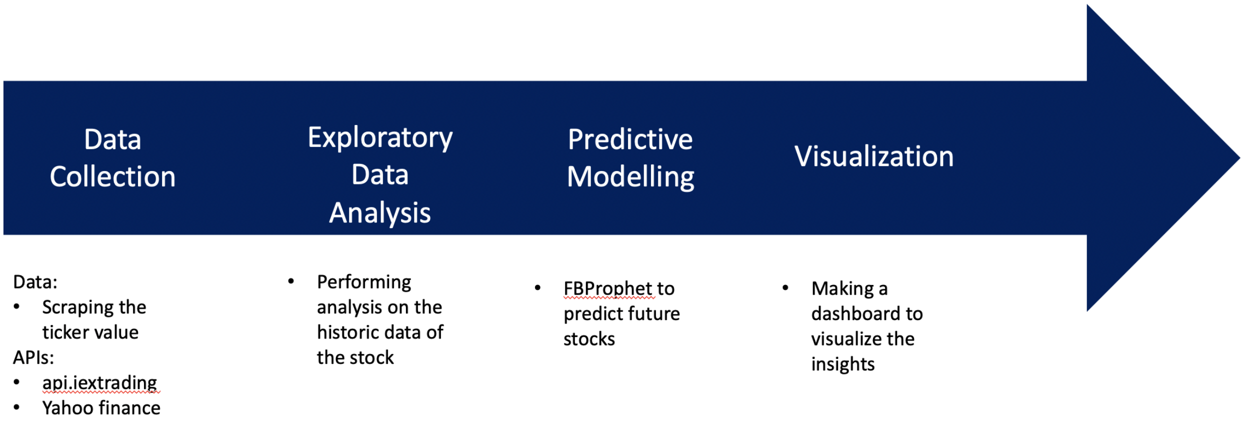
3.1. KPI 1 - Market Sentiment Analysis

In order to get a sense of the running sentiment about a company among all of its stakeholders, we planned on performing sentiment analysis on the news published about all the financial firms. Here is the Data Science Pipeline for KPI-1:


Data Collection
Data Cleaning
Sentiment Analysis
Visualization
Data:
• Website: thedailymail.com
Tools:
• BeautifulSoup
• Scrapy
• Fortune500 companies
• Shortlisting articles with company news in it.
• Finding the polarity values of the news articles
• Making a dashboard to visualize the insights

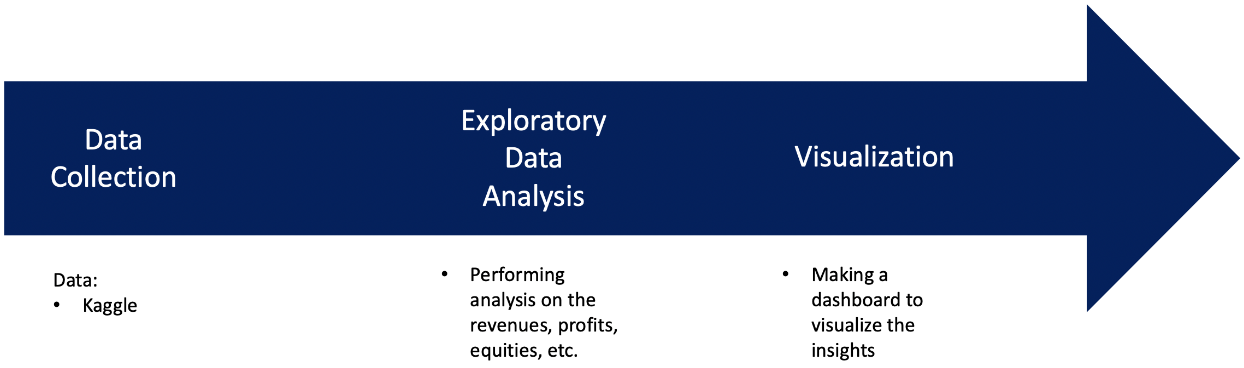
3.2. KP1 2 - Stock Prediction

The stock progress of a firm is one of the most common factors that's considered while the key investor makes the decision of choosing a company to invest. We pulled the historic stock information for all the firms over the past years from Yahoo finance and an api called iextrading. Here is the Data Science Pipeline for KPI-2:


Data Collection
Exploratory Data Analysis
Predictive Modelling
Visualization
Data:
Scraping the ticker value
APIs:
api.iextrading
Yahoo finance
Performing analysis on the historic data of the stock
FBProphet to predict future stocks
Making a dashboard to visualize the insights

3.3. KPI 3 - Sector-wise Analysis and Financial KPIs

Analysis on not only the progress of a company but also the sector that it belongs to can give us deeper insights about the companies. We can leverage this information to judge firms that can take a hike or undergo a financial dip. Here is the Data Science Pipeline for KPI-3:


Data Collection
Exploratory Data Analysis
Visualization
Data:
Kaggle
Performing analysis on the revenues, profits, equities, etc.
Making a dashboard to visualize the insights

4. Overview of the KPIs

4.1. KPI 1 - Market Sentiment Analysis

In order to get a sense of the running sentiment about a company among all of its stakeholders, we planned on performing sentiment analysis on the news published about all the financial firms. Naturally, the companies having a positive polarity sentiment value was a good choice for investment and vice-versa. The aim behind including this KPI was taking a more informal parameter in combination with other finance based KPIs to make a more solid decision.

We used **TextBlob** to perform sentiment analysis on the news features. Sentiment analysis is basically the process of determining the attitude or the emotion of the writer, i.e., whether it is positive or negative or neutral.

Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms.

The main types of algorithms used include:

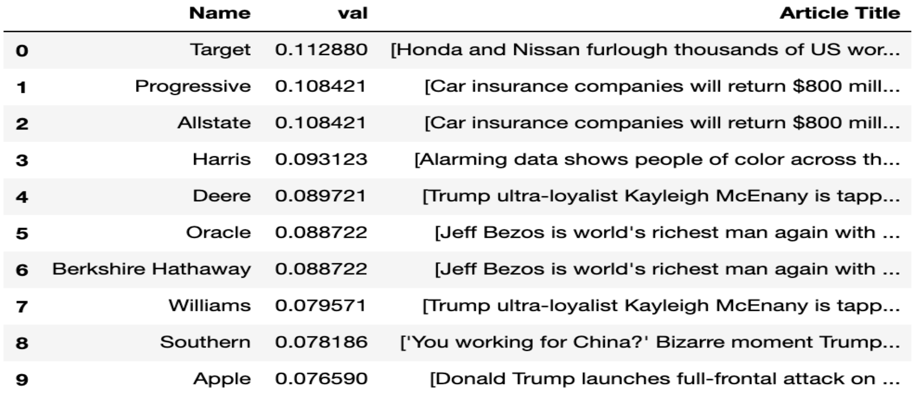
- Rule-based systems that perform sentiment analysis based on a set of manually crafted rules.
- Automatic systems that rely on machine learning techniques to learn from data.
- Hybrid systems that combine both rule-based and automatic approaches.

TextBlob[3] is a Python library that is built on top of nltk. It's easier to use and provides some additional functionality, such as rules-based sentiment scores. The *sentiment* function of textblob returns two properties, **polarity**, and **subjectivity**.

Polarity is a float value which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

Workflow:

- Scraping news articles all over the world from the website https://www.dailymail.co.uk [4]. The Daily Mail is a British daily middle-market newspaper published in London in a tabloid format. Founded in 1896, it is the United Kingdom's third-highest-circulation daily newspaper, after Metro and The Sun.
- Shortlisting the articles that have information about the companies that we are working on (Fortune 500).
- Perform sentiment analysis on articles corresponding to every company to get polarity value (-1 to +1 indicating negative to positive content). The output dataframe looked as follows:



	Name	val	Article Title
0	Target	0.112880	[Honda and Nissan furlough thousands of US wor...
1	Progressive	0.108421	[Car insurance companies will return $800 mill...
2	Allstate	0.108421	[Car insurance companies will return $800 mill...
3	Harris	0.093123	[Alarming data shows people of color across th...
4	Deere	0.089721	[Trump ultra-loyalist Kayleigh McEnany is tapp...
5	Oracle	0.088722	[Jeff Bezos is world's richest man again with ...
6	Berkshire Hathaway	0.088722	[Jeff Bezos is world's richest man again with ...
7	Williams	0.079571	[Trump ultra-loyalist Kayleigh McEnany is tapp...
8	Southern	0.078186	['You working for China?' Bizarre moment Trump...
9	Apple	0.076590	[Donald Trump launches full-frontal attack on ...

4.2. KPI 2 - Stock Prediction

The stock progress of a firm is one of the most common factors that's considered while the key investor makes the decision of choosing a company to invest. We pulled the historic stock information for all the firms over the past years from Yahoo finance and an api called iextrading.

Yahoo Finance is a media property that provides various financial news and data including stock details, quotes, press releases, financial reports. The IEX API or iextrading is a set of services offered by The Investors Exchange (IEX) to provide access to data from the Exchange to developers and engineers for free.

However, just making a decision on the past stock value wasn't enough. We thus predicted the future stock for these companies for the upcoming months using the fbprophet model.

Working of the prophet forecasting model:

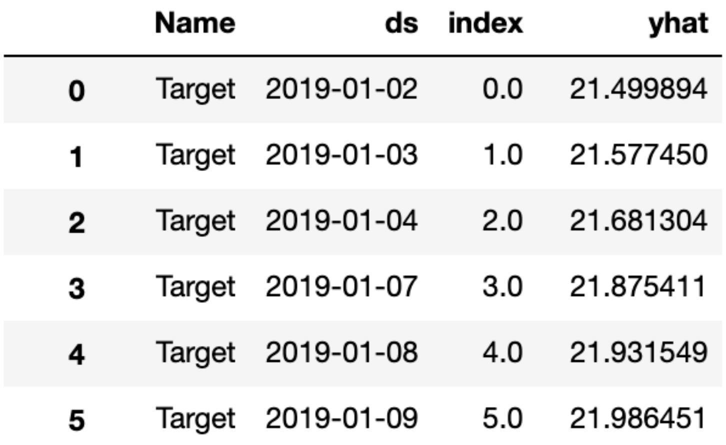
The Prophet uses a decomposable time series model with three main model components: trend, seasonality, and holidays [2]. They are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t$$

- g(t): piecewise linear or logistic growth curve for modeling non-periodic changes in time series
- s(t): periodic changes (e.g. weekly/yearly seasonality)
- h(t): effects of holidays (user provided) with irregular schedules
- εt: error term accounts for any unusual changes not accommodated by the model
- Using time as a regressor, Prophet is trying to fit several linear and nonlinear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in Holt-Winters technique . Prophet is framing the forecasting problem as a curve-fitting exercise rather than looking explicitly at the time based dependence of each observation within a time series.

Workflow:

- Scraping the ticker value for each firm (as needed for stock data analysis) using 'https://api.iextrading.com/1.0/ref-data/symbols' API
- Get the stock data on the decided window to monitor the performance using yahoo finance.
- Predict future stock using **fbprophet**. The output dataframe looks as follows:



	Name	ds	index	yhat
0	Target	2019-01-02	0.0	21.499894
1	Target	2019-01-03	1.0	21.577450
2	Target	2019-01-04	2.0	21.681304
3	Target	2019-01-07	3.0	21.875411
4	Target	2019-01-08	4.0	21.931549
5	Target	2019-01-09	5.0	21.986451

4.3. KPI 3 - Sector-wise Analysis and Financial KPIs

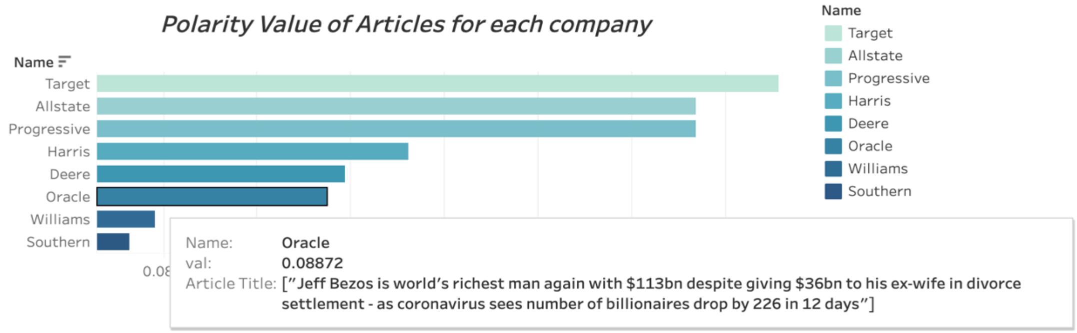
Analysis on not only the progress of a company but also the sector that it belongs to can give us deeper insights about the companies. We can leverage this information to judge firms that can take a hike or undergo a financial dip. The most relevant example for this use case is the fact that the current outbreak has led a sudden hike in all the retailing business and IT industry firms that facilitate work from home. Predicting this behavior can help us make the right investment decision.

Workflow:

- Performed comparative analysis on Sector and industry
- Performed analysis on monetary KPIs such as assets, profit, total equity, revenue, etc.

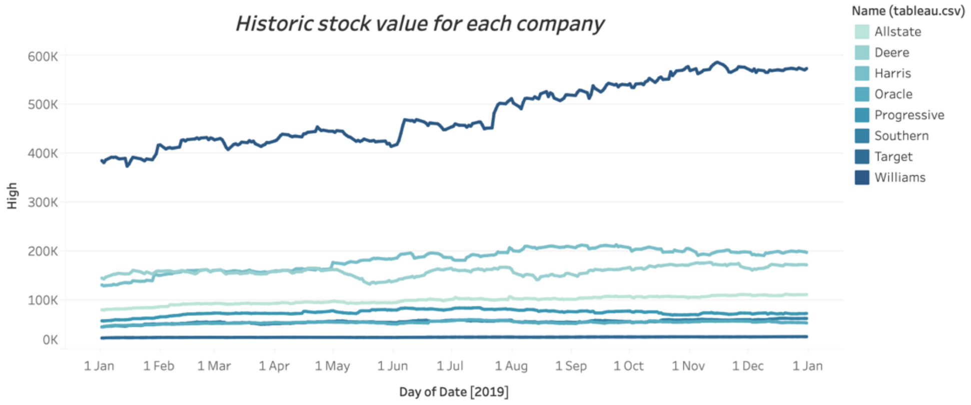
5. Data Product

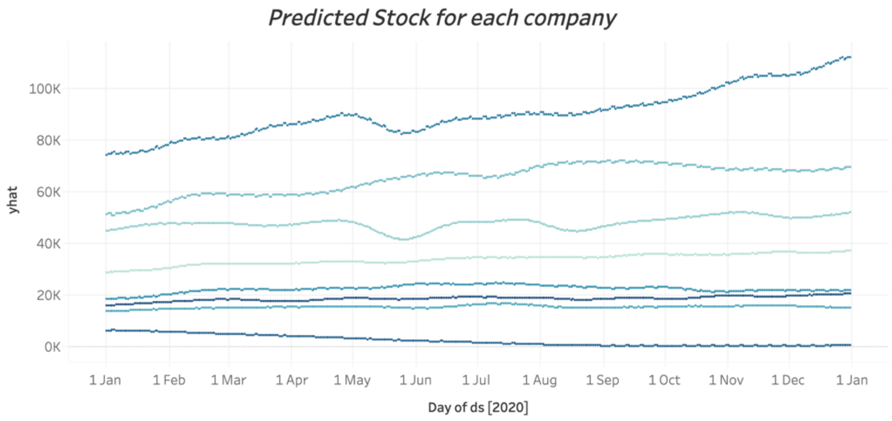
5.1. KPI 1 - Market Sentiment Analysis



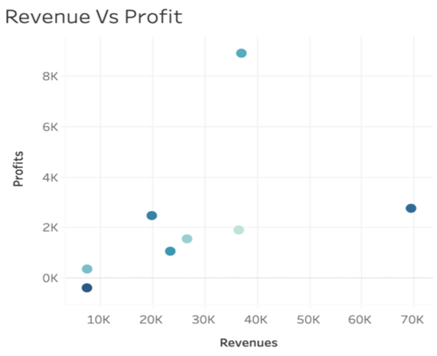
Name: Oracle
val: 0.08872
Article Title: ["Jeff Bezos is world's richest man again with $113bn despite giving $36bn to his ex-wife in divorce settlement - as coronavirus sees number of billionaires drop by 226 in 12 days"]

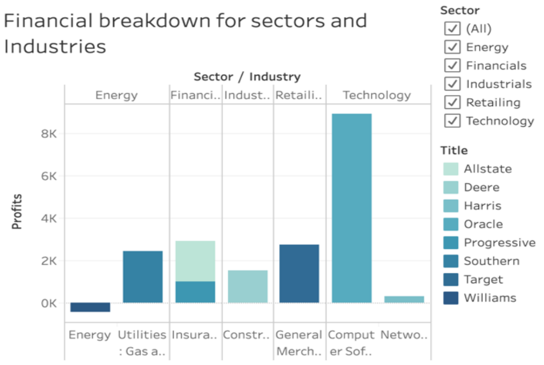
5.2. KPI 2 - Stock Prediction

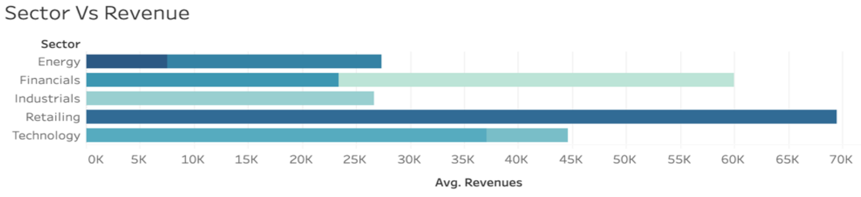




5.3. KPI 3 - Sector-wise Analysis and Financial KPIs







6. Technology Stack

EDA	Numpy, Pandas, Dataprep.eda
Scraping	BeautifulSoup, dailymail, iextrading, yfinance
Sentiment Analysis	TextBlob
Prediction	Fbprophet
Visualization	Plotly, Tableau

7. Evaluation

As mentioned earlier, the data product suggested the right sectors and industries to invest-in, given the current financial situation in the industry.

Also, the data product not only considers all the finance based factors but also other informal ones like market sentiment which denotes the all-round success rate of a firm.

8. Lessons Learnt

This project served as a great means to implement all the technologies learned in the CMPT 733 course. We implemented a combination of a lot of technologies to address various parts of our project. First and foremost challenge was to come up with a good problem statement and a workflow and pipeline to address the problems part by part.

In the ETL process, we learned various ways of extracting data by scraping it from different websites and cleaning it to work with it. For modeling, we learned about Facebook's prophet model which helped us make stock predictions. It was also a challenge to create visualizations that will be understood by people from other domains such as business backgrounds. We learned about visualization principles and tried to implement them in the best way possible. Finally, it was a different experience presenting the project virtually instead of the traditional ways. Overall, it was a great learning experience cumulatively.

9. Challenges and drawbacks

a. Selecting the right KPI:
   The use of a certain KPI might be relevant to one company and completely irrelevant to another depending on the company's business model and the stage the company is at. For example, consider Profit Increase, if a certain start-up shows a profit change of around 4-5% in its initial years, it actually shows a very good progress and investing in this company might be a good bet. At the same time, if a well established firm shows the same numbers, it actually means that the growth is not at its peak for this company.

b. Restrictions on the amount of information that can be scraped using a particular APIs:
   A lot of APIs only provide a certain amount of information from a free account. So to increase the scope of the project, we might need to establish a paid account with APIs like yfinance for uninterrupted service.

c. Limitations of Fbprophet:
   Fbprophet works quite well when it's fed with only one attribute that corresponds to the historic time frame to predict future values. In our case we fed the model with the highest stock value for a company for a particular day. However, it is incapable of predicting data of multiple dimensions. We overcome this shortcoming by involving an additional regressor using an add_regressor() function [9].

d. Summarising the news feature:
   Initially we only displayed the sentiment values for every firm in the final data product. However, it was unclear to a viewer, the reasons behind the model claiming the negative or positive values. We thus decided to include an additional feature. The final data product includes summarized views of articles that led the polarity values to what they are displayed. This includes the viewers to not only know the sentiment value but also get a gist of the articles.

e. Unpredictable market behaviour:
   Our model has a basis that companies with negative sentiment value might show a decline in performance and thus it is considered as a bad move to be investing in the firm. However, it is said that any publicity is good publicity. For example, Cybertruck had very bad reviews after the major failure in the launch session. However, the company Tesla did not face any backlash and this in fact led more and more people talking about it and the sales increased.

10. Future Work

a. Net Promoter Score (NPS):

Our data product aims on making investment decisions based on several KPIs, both financial and informal. We intend on adding as many KPIs as possible to make our predictions more solid. However, a lot of KPIs involve information that has sensitive data and generally this information isn't available on any public domains. This information limits us to include only budget-linked KPIs. While budget-linked KPIs are important, the ultimate indicator of a company's potential for long-term success is in its Customer Satisfaction quantification. The Net Promoter Score (NPS) [6] is the result of calculating the various levels of positive response that customers provide on very brief customer satisfaction surveys. The NPS a simple and accurate measurement of likely rates of customer retention (future sales to current customers) across your revenue base, and of potential for generating referral business to grow that base.

b. Fiscal KPIs:

The viability of a business model can be accurately judged using the following fiscal metrics[7]. We intend on including most of these to make an improved version of our model. Some of the metrics that we have shortlisted are as follows:

- Earnings before interest and taxes (EBIT)
- Economic value added (EVA)
- Berry ratio
- Contribution margin
- Liquidity ratio

c. Context Mining:

After scraping the entire news corpus from daily mail journal, we needed to shortlist all the articles that included any information about financial firms. However, a normal python parser or regex string matching has drawbacks. Often times, the name of a company can have multiple meanings. For example, while trying to fetch the data for Apple Inc., it's highly likely that a news about any local farms about apples might get shortlisted too. In order to avoid that, we intend to use context mining to shortlist the correct news.

11. Conclusion

Investing smart is one of the most key decisions that every broker and investor has to make. Making these smart choices is very crucial and we hope the project we've built can be used for these decisions. Our project aims to evaluate primary metrics that can help decide the right company to invest in. This crucial decision was made based on market sentiment analysis, stock values, sector wise analysis and financial KPIs like profit, revenue, total equities. Including sentiment based KPI helped us make the model based on not only non-budgeted performance indicators but also on informal factors for accurate insights. It is exciting to see data science expanding its lengths and breadths beyond the IT industry and having use cases in domains like business and finance. In conclusion, our model assisted investors to make smart decisions by deriving intuitive results leveraging analytical and prediction models.

References

1. https://sixfigurestockportfolio.com/
2. https://towardsdatascience.com/a-quick-start-of-time-series-forecasting-with-a-practical-example-using-fb-prophet-31c4447a2274
3. https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/
4. https://www.dailymail.co.uk/home/index.html
5. https://blog.api.rakuten.net/api-tutorial-yahoo-finance/
6. https://www.accountingdepartment.com/blog/12-key-performance-indicators-you-should-be-tracking
7. https://www.klipfolio.com/resources/kpi-examples/financial
8. https://www.business2community.com/finance/8-financial-kpis-to-help-improve-performance-02242040
9. https://towardsdatascience.com/time-series-prediction-using-prophet-in-python-35d65f626236