**DATA MANAGEMENT PLAN**

[1 Data Sources Used](#_b072co8ndp54) 2

[2 Data Usage](#_3xbmlc28qifd) 2

[3 Data Preparation](#_3ym0pry5ic5u) 3

[4 Data Storage](#_9b3ifng0ss9p) 3

# 1 Data Sets Overview

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Source** | **Format used** | **Format to be used** | **Update Frequency at Source** | **Granularity** | **Last Updated** |
| 1 | [BSE](https://www.bseindia.com/market_data_products.html?flag=real) | Data - wrangled CSV | API | Daily | High | 19/11/2024 |
| 2 | [Aswath Damodarum - NYU Stern](https://pages.stern.nyu.edu/~adamodar/) | Open sourced CSV | Automated downloads | Monthly | High | 04/01/2024 |
| 2 | [Reddit](https://www.reddit.com/dev/api/) | Open sourced Kaggle CSV | API | Daily | Medium – Finance communities | 14/12/2022 |
| 3 | [Twitter](https://docs.x.com/x-api/posts/search/introduction) | NA | API | Daily | Medium - Finance communities | NA |
| 4 | [Youtube](https://developers.google.com/youtube/v3/docs/search/list) | NA | API | Daily | Medium - FinFluencers | NA |
| 5 | [Economic times](https://buy.indiatimes.com/ET/not/plans?marketingConfig=9&acqSource=Google_Paid_Marketing&acqSubSource=Remarketing&utm_source=Google&utm_medium=cpc&utm_campaign=ET-Google-Website-Search-RLSA-India-CPC-Purchase-Nov24_CG&utm_id=21894038280&utm_term=remkt&utm_content=Brand&gad_source=1&gclid=CjwKCAjwktO_BhBrEiwAV70jXqydw3qv07WZxPocOCSGIkY4voK9coqK3TBCaLgIL3Vky0xB6OLuIBoCjUYQAvD_BwE) | NA | e-paper | Daily | Low | NA |

2 Data Usage

**Stock Financial Metrics**: Analyze key financial indicators to understand how company fundamentals correlate with share price movements, utilizing conventional valuation methodologies.

**Company Announcements**: Interpret Management Discussion and Analysis (MD&A) sections of annual reports and assess the nature of corporate announcements to derive company-specific sentiment.

**Finance Communities on Social Media**: Assess the influence of online finance communities across platforms on overall market sentiment, with a particular focus on identifying manipulation tactics such as pump-and-dump schemes.

**Finfluencers**: Evaluate the impact of financial influencers on investor behavior, emphasizing the detection of manipulative strategies to help protect retail investors from misinformation and financial losses.

**News Media**: Monitor news mentions to identify media-driven hype and extract deeper industry insights

## 3 Data Preparation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Intended Use** | **Transformation** | **Transformation code file** | **Columns Retained** |
| bse\_modelling6.csv | This data is crucial to perform valuation and present numbers for the machine learning model. | A stock research website was web scraped , furthermore necessary changes to ensure data consistency | making\_bse.  Sorting\_bse,  **Modelling.csv** | ALL |
| Industry\_rates.xlsx, interest\_coverage.csv | Critical to the valuation feature for industry numbers | Modelling.csv columns made are added to check veracity of this data | **Sorting\_bse.py** | Industry\_name, Number\_of\_firms, reinvestment\_rate, expected\_growth\_EBIT, average\_operating\_margin |
| posts.csv | Used for the sentiment analyser feature | Data from various sub reddits cogulated | **Sentiment.R** | User, title, text, post\_date, upvote, comments |

## 4 Data Storage

* Data is stored on an **MySQL server**
* A **MySQL database** is used to access dataset FinancialData
* The database was created by defining a schema that somply just lists all columns in text datatype.
* Datasets were ingested into the MySQL database from CSV files
* The **website** is connected to the database via a **secure API**, allowing real-time data retrieval for model training.