Monitoring Market Volume and Generating alerts using Sentimental Analysis of financial news articles for Risk Evaluation

Likitha P[1] ,Sonal S R[2] , Dr.G.S.Mamatha[3], Rekha B S[4]
1, 2(Students, Information Science and Engineering,
R V College of Engineering)
3, 4 (Professors, Information Science and Engineering,
R V College of Engineering)

Abstract : Digital platforms are providing a huge amount of data for analysis. And such analysis has become the need of the hour in various fields. Stock markets is one such field, and financial institutions handling the data related to real time deals will benefit if a system can predict the number of deals that can take place and thus gauge the storage requirements prior to the situation of risk. This helps the organization to avoid data storage breaches that might arise due to storage capacity exhaustion. This paper presents a design to analyse market volume in terms of liquidity of transactions by quantifying the impact of news on the financial market

Keywords* — *Volume, liquidity, sentiment analysis, NLP, classification, Bag of words, grafana , influx

I. Introduction

Market volume is the total number of transactions happening in a period of time, it can be transactions of equities, commodities or currencies .Market Liquidity refers to the ease with which assets can be sold or bought and readily converted to cash at a stable intrinsic price without affecting its market value. As noticed from time to time, the market is affected by a lot of factors which define the market price, market volume, liquidity etc. As it might also affect the economic condition of the country as well, it becomes important to analyse the importance of different factors affecting the market. The reason for stock volatility can be due to aggregated behaviour of stockholders, influencers, daily news, financial announcements, budget distribution, tweets and many more. Market volume and liquidity are interrelated terms, which affect each other on a daily basis, like higher trade volume indicates higher demand of the stock in the market .This in turn increases the transaction creating high data inflow of transaction details.

II. Motivation

The liquidity of the market decides the volume of transactions taking place which is in accordance with the amount of storage required by the financial institutions to manage or keep track of stock deal records for further procedures. Many times huge changes in the market may cause the data storage breach in case of limited storage conditions. Hence there is a need to analyse the risk in advance by considering news affecting the market liquidity, to notify the upcoming storage needs and take proper measure of scaling accordingly. maintaining the integrity of the specifications.

III. Litrature Survey

The data model consists of 3 models which mainly studies the impact of news on financial markets in terms of daily mean cumulative returns, frequency measure, liquidity, price fluctuations. The Sentiment model handles news events and their corresponding analytics data. Market model handles the prediction of market measures such as liquidity, price and volume. And the Comparison model links sentiment and market models defining a set of parameters that can drive the impact of news of financial markets. It shows that negative news has a larger impact on the market compared to positive news. The negative market news decreases market liquidity and positive news increases the liquidity [1].

The information derived from news is recently used in predicting the direction of stock movement as a classification task and precise prediction of stock price as a regression problem. This paper presents the impact of news in asset volatility and statistics prediction rather than price. Analysis is performed using stocks in the US market. Machine learning models such as Latent Dirichlet Allocation is used to extract useful information from news and a simple naive Bayes classifier for direction of movement of stocks [2].

Sentimental Analysis is applied on the news data for the interpretation and extraction of emotions from different sources. The need for this has increased with the rise in social media and huge data availability. It proposes a deep learning model to be most fit since it has the ability to analyze the great amount of data to be analyzed by NLP for context and gramatics. News headlines are used for basic linguistic preprocessing such as (removal of stopwords and special characters, lowercasing).The impact of positive statements is ignored since it is not clear that positive statements always cause the market to grow; it may be the other way round as well. This helps the model to learn freely by itself [3].

Liquidity prediction is a difficult task since it is highly volatile and dynamic in nature. It defines a fully connected neural network composed of Multilayer Perceptron(MLP), Mixed Deep Learning(MDL) and linear Regression. Various metrics such as Mean absolute error and mean squared error are continuously monitored to improve the model. Among these, LR proved to be the worst model with highest mean squared error and mean average error. Deep learning models proved to be the best technique for liquidity prediction [4].

It suggests that the public mood is a correlated and predictive economic indicator. Daily twitter feed is analysed using two public mood tracking tools Opinion Finder and Google Profile of Mood States.
Opinion Finder predicts the mood as positive or negative whereas Google Profile of Mood States verifies the mood in 6 dimensions as Alert, Calm, Happy , Kind , Sure , Vital . These parameters are predictive of Dow Jones Industrial Average(DJIA) closing values[5].

The News Sentiment is Analysed using Thomas Reuters News Analytics tool (TRNA) which returns a single vector which is further used in the calculations of popularly known volatility models such as GARCH(Generalized Autoregressive Conditional Heteroscedasticity), GJR and EGARCH (Exponential GARCH) using DJIA(Dow Jones Industrial Average Dataset) .Vendors such as TRNA and Raven pack provide sentiment scores to provide direct indicator to traders about the predicted market changes. The text mining tools employ pattern recognition methods to analyse words, patterns, and the originality and relevance of the news items to a sector. These news items are converted into quantifiable sentiment scores at a sentence level. Any new item in the model is tagged with an exact timestamp and the list of topics it mentions and the companies that can get affected with the news of concern. Five main categories considered are Relevance of news item to asset, Sentiment (Negative, positive or Neutral), Novelty(previously dealt with or not), Volume(count) and Headline Classification (specific analysis)[6].

The paper deals with NSIA(News Sentiment Impact Analysis) Framework which has a model that mainly considers three types of parameters -financial, sentiments and evaluation metrics. It also proposes a GUI based system to make it user friendly. The data is collected from various corpuses, using API calls, rss feeds etc, and result obtained is usually in the form of Json format. Different preprocessing techniques are applied. It talks about the calculation of sentiment score using a knowledge based method using the context of words, ML techniques using a pre-trained classifier, NLP techniques and Thomas Reuters framework which calculates sentiment score of the news provided by its clients. Various terms related to finance are discussed accordingly[9]

The impact of news sentiment on companies is studied by considering 87 companies belonging to different sectors which are widely reported by Reuters. The data is collected for a period of 7 years. A network is maintained to study the impact of new sentiment on a company and the organisations associated with it, as part of volatility or stock price .Named Entity Recognition is performed for the identification of entities of type ‘organisation’ using convolutional neural network(CNN) and LSTM model. Each organisation can have different types of references like ABC Org,ABC Ltd etc, measures are taken to obtain a common word for each of them. Clustering technique is performed to form a network. The analysis is performed at sentence level which takes context of news into consideration and then assigns values -1,1 to the entities identified with -1 being negative impact and 1 being positive impact. Grouped sentiments are also applied in relation to study of market data[10]

The changes in RSS news feeds is particularly studied for its impact on the stock market prices. The data for stock and news feeds is collected from Amman Stock Exchange for company ARBK. A lot of preprocessing is done to remove duplicate and improperly formatted data. The cleaned sentence is then parsed. The emotion of news on a particular topic is analysed by using a NLP module. Parts of Speech(POS) tagging is carried out. An algorithm is then defined to get the sentiment score which inturn decides the positive,neural or negative impact at sentence level. Various methods C4.5, moving average and ID3 are used to study the fluctuations in stock.
Also proposes a future plan of adding some stock level indicators along with news feed for better accuracy[11]
The news data is applied with various levels of POS tagging to get the phrases. Only “Excellent” and “poor” are used as indicators. The seed set method is used to extract the features, which also makes use of manual intelligence as only considering adjectives might not be effective in finding good phrases. The importance of noun-verb combination in determining the context or meaning of a sentence is explained with suitable examples. Model is based on the proposed idea which is used to classify the financial data into suitable predefined class [12]

Introduces a method for sentiment analysis of news affecting financial markets using lexical resources and associations of words. The news data is collected for a period of about 14 years from 2000 to 2014. Proposes the idea of labelling the sentence with emotions along with positive or negative impact. News data related to financial content consisting of 918,427 documents. Various Natural language processing operations such as tokenization, stop word removal are applied as per requirement in further stages. Algorithms defined to get the Sentiment Scores work at document, sentiment or word level. This paper introduces the application built using the model built in 3 modules, a module to accept the news data as a word editor, second module to present stock quotations of different companies to users, third module presenting visualization of the data [13]

Studies the impact of COVID-19 news on the stock data .The data is collected for a period of about 6 months from January to June 2020, including all the articles that were published in three news platforms. It tries to establish a correlation between market value and news sentiment score. The COVID news data is first tokenized, Then Bidirectional Encoder Representations from Transformers(BERT) method is used to get the Sentiment score. The variance in sentiment from one day to another, also the volume of articles obtained serve as the attributes for the model. Volatility and growth index are metrics to analyse the direction of movement of stocks. [14]

The impact of news is studied by collecting the data for the pharmaceutical market. Data is collected for a period of 6 months for each company using the Beautiful Soup module of python. Analysis is also carried out on a quarter basis if the news data consists of Q1, Q2 in their content. A python module called “pattern” is made use to generate n gram vectors from news statements. The method started by first considering unigram works, later it was proven that context of the word is very important to determine the impact , hence it was improvised to bigram, trigram , n -gram etc [16]

Hence the study of papers relevant to the topic “Sentimental Analysis of News impacting financial Markets” helped to develop domain knowledge to go ahead with monitoring the market volume and certainly carrying out Risk Analysis of in financial systems.

IV. Problem Statement

When the market is at high liquidity level, there is a huge amount of data inflow every day into system, the insert counts varies in range of billions each day This inflow of data leads to a chain of reaction like data storage breach causing bottleneck in steaming of data in and out .This also leads to exhaustion of resources causing latency. .So monitoring this inflow of data and alerting on unusual trends becomes very important in order to maintain data storage and resource usage.

V. Objective

This paper focuses on 3 major aspects:

- Effective Sentimental Analysis of Real Time News affecting market behaviour
- Envisioning Market Volume for efficient utilization of data storage and resources
- Visual Monitoring of data inflow
- RealTime alerts on critical issues

VI. Methodology

To achieve the stated objectives these are the steps or methods necessary to implement the automated monitoring tool:

1. Setting up schema based on the use case
2. Aggregation of data from multiple sources
3. Creating APIs to perform preprocess and necessary calculations , transformation to generate the data required
4. Setup automated jobs to perform the above steps regularly in real time
5. Establishing kafka connection to stream data
6. Publish data to Influx db within the constraints of the time series database.
7. Creating an interactive dashboard with multiple panels shows analytical visualization for the results
8. Setting up alerts based on analytical patterns , rules based and sentiment analysis.

News Sentiment Analysis is essentially a text classification task which can be categorized into the following steps:(as shown in Figure 1)

(i) Collection of News dataset for training.

(ii) Extraction of keywords and structuring the data

(iii)Creating the Bag of words

(iv)Vectoring the dataset

(v) Using suitable model to performing scoring the sentiment

Real time news is collected from data sources which has two main parts: headlines and news contents. More importance can be given to the information in the headlines than to news paragraphs. The data collected is such that it contains news headlines of the day and the corresponding attribute representing changes in stock market direction such as high, open and close quotations of stock of the same day. This data can be collected for the instruments of concern in this study.

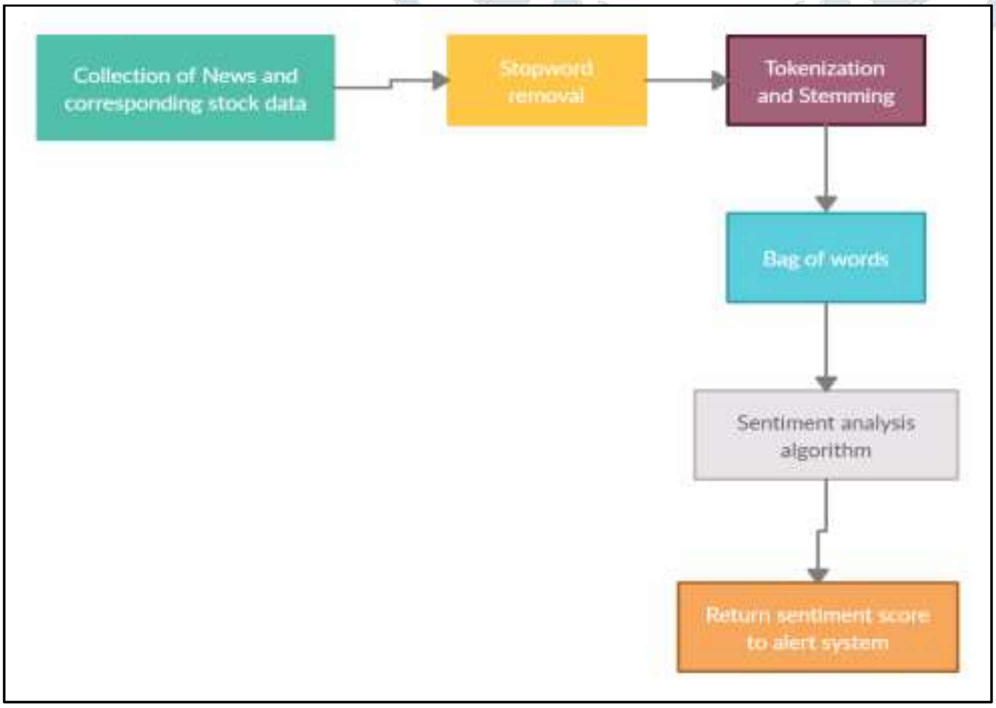

Collection of News and corresponding stock data
Stopword removal
Tokenization and Stemming
Bag of words
Sentiment analysis algorithm
Return sentiment score to alert system

Figure 1:Sentiment Analysis of News

Preprocessing techniques should be applied to remove any kind of noise involved in the data collected before applying NLP techniques. As the data collected is in the form of natural language which can have words of different cases(Upper case, lower case, camel/ pascal case etc), each word will be treated differently while building a model, hence all the words are converted into a common case(preferably lower case) .

Also any special characters such as comma, apostrophe, colon, semicolon etc can be removed. Also various other preprocessing technique can be applied as per requirement such as:
Stop word removal- These words are just helping words and do not provide much meaning such as a, an, the etc

Stemming - To get the root word from different usages by removal of prefixes and suffixes. If the data is obtained from different news articles, they can be combined to represent a single news entity for all future processing Tokenization -To split the sentence into individual tokens

Bag of Words Model is a method applied to natural language data to get the vocabulary of words present and their corresponding distribution . Bag of words is applied to the news data using CountVectorizer from the feature extraction module of sklearn.

The model is trained using the vectorised data.

Vectorization is the process of mapping a word or phrase with a real number usually used for word predictions and semantics.

The sentiment scoring can be done using [6]TRNA(Thomas Reuters News Analytics) Tool, which returns a single vector representing various sentiment attributes such as novelty(whether the news is old or new), relevance(to the specific instrument, sector) , volume(count of news data on similar topic). The sentiment score is given as 0(neutral), 1(positive impact on instrument and sector), -1(negative impact on instrument and sector) as shown in Figure 2.

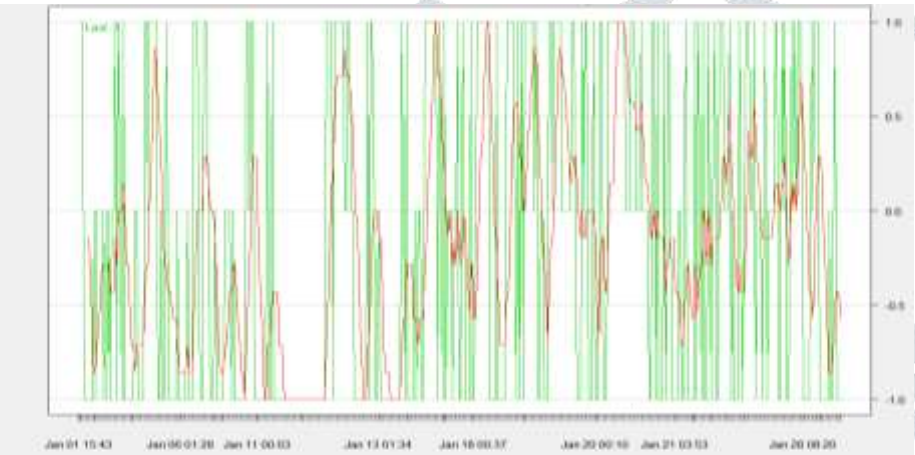


Figure 2. TRNA sentiment scoring

Sentiment scoring can be done at document , sentence or window level. The bag of words model is used accordingly.

At a document level, the scores of all keywords are added. Similarly for sentence level, only keywords from chosen sentences can be considered, whereas in widow level a certain number of words to the left and right of the Instrument Keyword is considered.

The sentiment score thus obtained is sent to the alert system to meet the requirements.

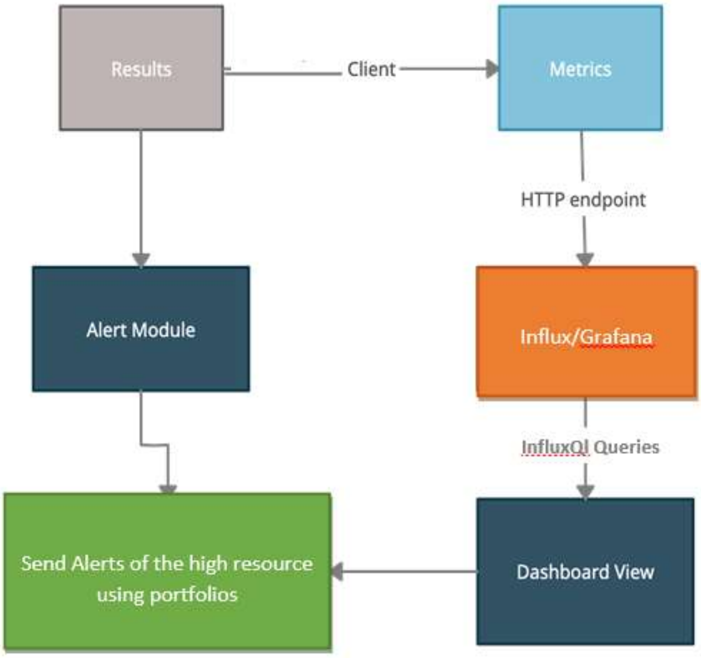


Figure 3. Visualization of exposed metrics

VII. Implementation And Results

Visualizing the metrics and stats aggregated from the current and historic data storage in the system is an important part of monitoring the data inflow trend and analyse the patterns. It is always a known fact that visual display of the results helps in analysing the data better than the static numeric output.

Grafana is an open source analytical visualization software , used to build this monitoring system . It has a variety of charts which are tightly bound with timestamp as a primary key to visualize the data .

The deal volume is aggregated based on the instrument , sector and region . This data is tightly bound with the timestamp and published to influxdb using apache kafka for data streaming . The data aggregation is done using complex querying like

“Insert into deal_stats select
sum(distinct deal_counts) as deal count ,
date as marker_date,
location as location,
group|-|location|-|marker as portfolio,
sum(distinct book_count) as book count ,
id as deal_id, instrument as deal_type
from deal_table a , marker_table b where a.id=b.id
where table_name =’table_name’ and
deal_id > (select max(marker_id) from deal_stats table)
group by 1,2,3,4”

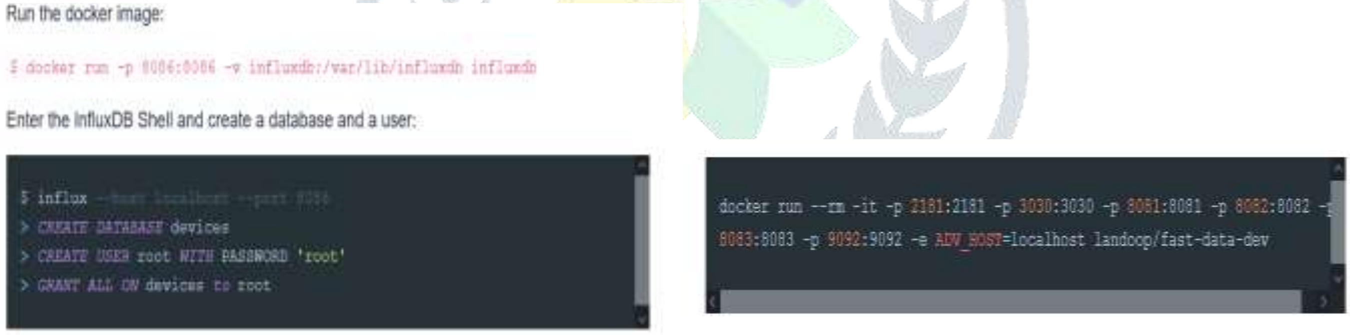
union

“Insert date_truc(hour,greatest(armrow_ts1,armrowts_2,armts3)), count(district deal) from table name a , marker name b where a.ts=b.ts groupby 1,2”

These queries are automated to run at equal intervals of time every day to collect the required data minimizing human manual time. Apache kafka is used for data streaming to publish the data into Influxdb. Influxdb is an open source time series database , which has high and fast availability of storage and retrieval of time series data.

The steps to connect kafka to influx db and publish data :(as shown in Figure 4 and Figure 5)

1. Setting up a docker development environment
2. Run the Kafka Influxdb Sink Connector
3. Set up producer to publish data to right topic
4. Consumer instance need to subscribe the topic



Run the docker image:

```
$ docker run -p 8086:8086 -v influxdb:/var/lib/influxdb influxdb
```

Enter the InfluxDB Shell and create a database and a user:

```
$ influx --host localhost --port 8086
> CREATE DATABASE devices
> CREATE USER root WITH PASSWORD 'root'
> GRANT ALL ON devices to root
```

```
docker run --rm -it -p 2181:2181 -p 3030:3030 -p 8081:8081 -p 8082:8082 -p
8083:8083 -p 9092:9092 -e ADV_HOST=localhost landoop/fast-data-dev
```

Figure 4. Run docker image

Influxdb is a time series database , where each timestamp has a data point , like discrete metrics. The tables in it are called ‘Measurements’ ,they store the key-value pair college ‘tags’ which are used for indexing and more preferable to query system and ‘fields’ are the actual data point value which are usually numerical and non indexed .

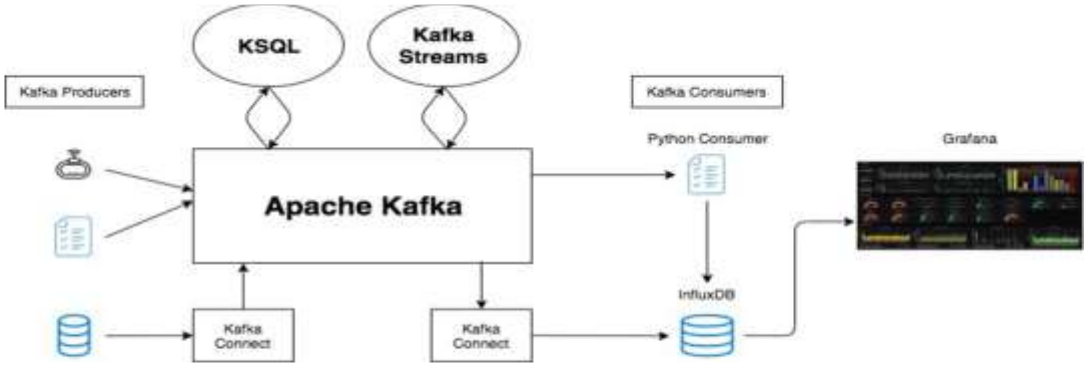


Figure 5. Flow of data publishing to influx to grafana

There are some internal checks and calculations done before the aggregated data is published to influx . The internal checks like day of week, month ends ,avrg deal , percentage change , new deal, addition of books and size used are done using a set of APIs created before publishing the data.

The setting up of tags and fields for the measurements in influx for the given data is done based on the use case .

Example: If the use case is to publish the distinct deals happened in a day , which even needs to be viewed with day of the week check .

measurement_name :{
field: {distinct deal count}
tag:{portfolio: 'name', deal_id: 'id', region: 'location', date: 'date' , dayofweek: 'day' ...}
ts: timestamp
}

Figure 6 is the grafana panel showing a line graph for 3 different instruments deal counts , the below is mapped with each point based on the timestamp of the deal being inserted into the database. The x-axis has the time stamp , the y-axis towards the left side have the deal count values in millions and the right side has the instruments names. We can the see the trend of data inserts and analyse the pattern , like the huge data spike happening at a regular interval of time .It is seen the data inserts to db is almost 5-6times more than the usual days on these data spike days. This panel can be further queried easily using the dashboard created just by a click to get deal counts on a particular day of every week or only on month ends or for a particular instrument . The value visualized can also be changed based on a use case , with just a click like view distinct book count or average deal count or record count .

This Monitoring dashboards simply the analytics based on use case by just with a click , which earlier needed complex querying .

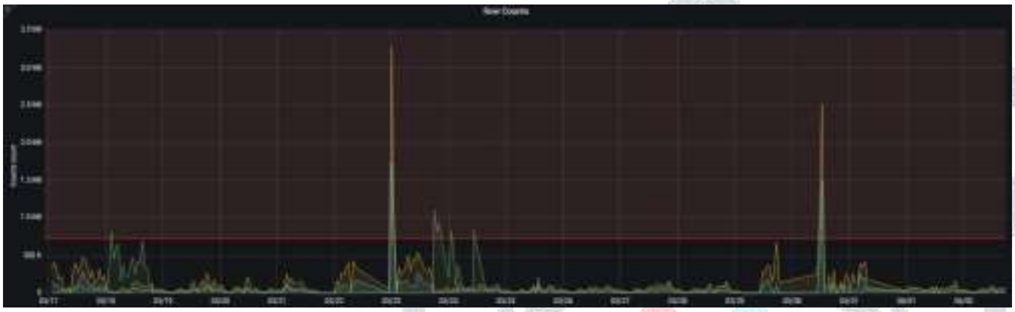


Figure 6. Panel of deal insert counts into database

Figure 7 shows the record count of each portfolio of given instruments . The x-axis is mapped to timestamp , the y-axis left side shows record value count in millions and right side shows portfolios. They even display a legend of a table showing the current value , average value for the given time period, min and max values of the selected time range .

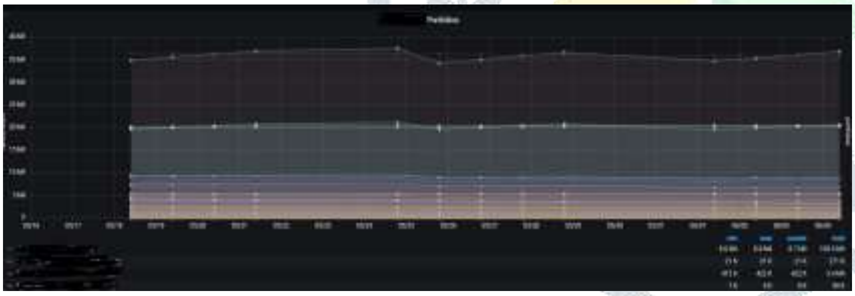


Figure 7. Visualization of Portfolio level breakdown of record count

Alerts are created for the following checks:

Fast Movers - The portfolios which have started trading more than 10% compared to last week

Outliers- The portfolios which have stayed dormant for an average period of time and suddenly start trading.

Book sneakers - Any additions of new books and instruments without pre notice

Market Volume predictors - The TRNA model is used to score the real time news sentiment on sectors and instruments. The score of 1 means a positive impact and high chances of increase in market volume of that sector instruments , 0 means neutral no much changes , -1 means negative impact which again indicates a bearish sentiment in the market and causes high market volume.

Based on all the above discussed methods, one can draw certain patterns for Risk Analysis. Context of news is necessary to determine the impact. Labelling technique is important to know the elements of news affecting the market volume and liquidity. Performing TRNA on the labelled dataset gives the count of positive, neutral and negative scoring.

With the help of these analysis and alerting, the future requirements of data storage and resources required for storing and processing the inflow of data respectively can be estimated. .This estimates an average amount of resources required. In case the resource requirement exceeds the available resources, an alert system will notify relevant support teams to expand the resource, storage capacity or to carry out a compression and archival of historical data to be moved to a cheaper storage container. This minimises the high risk of sudden un alerted data spike which would cause an uncertain chain of reactions like latency, storage breach.

VIII. Conclusion

The behaviour of market performance is mainly affected by the news involving the aggregated behaviour of stockholders, influencers, tweets, social media, budget, political statements etc. Based on the news sentiment, the volume of transactions in the market is affected. This paper deals with solving difficulties faced due to a sudden data spike or unusual data growth due to changes in market trends. This in turn links to chain of reaction causing system overhead, delaying the running jobs for over a period of time.

The future enhancements in the field of Risk Analysis can be suggesting Automated Archival period to the concerned authorities and minimizing the human intervention by building a smart system that is able to take required decisions. Example: A system should be able to analyse the suggested archival time and predicted volume and make decisions of scaling, storage and resources , archival of not so important data of a certain time range.

IX. Acknowledgment

We would like to mention our special thanks to our guides (Prof. Rekha BS and Dr. GS Mamatha) who have been a constant help in guiding us while working on this paper, which also helped us in doing a lot of research and helped us learn and explore many new things.

References

[1] I. A. Qudah and F. A. Rabhi, "Systematic Approach to Quantify Impact of News Sentiment on Financial Markets," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019.
[2] Adam Atkins, Mahesan Niranjan, Enrico Gerding, "Financial news predicts stock market volatility better than close price", The Journal of Finance and Data Science, Volume 4, Issue 2, 2018.
[3] Vicari, M., Gaspari, M., "Analysis of news sentiments using natural language processing and deep learning" . AI & Soc (2020).
[4] Pham Quoc Khang, Marcin Hernes, Katarzyna Kuziak, Artur Rot, Wiesława Gryncewicz, "Liquidity prediction on Vietnamese stock market using deep learning", Procedia Computer Science,Volume 176, 2020.
[5] Johan Bollen, Huina Mao, Xiaojun Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, Volume 2, Issue 1, 2011.
[6]David E. Allen, Michael McAleer,Abhay K. Singh, "Machine News and Volatility: The Dow Jones Industrial Average and the TRNA Sentiment Series", TI 2014-014/III,Tinbergen Institute Discussion Paper
[7] Ritu Yadav, A. Vinay Kumar, Ashwani Kumar,"News-based supervised sentiment analysis for prediction of futures buying behaviour",
IIMB Management Review, Volume 31, Issue 2,2019.
[8] Vayanos, Dimitri & Wang, Jiang, Market Liquidity - Theory and Empirical Evidence. Handbook of the Economics of Finance,2012
[9] Qudah, I., & Rabhi, F. A. News sentiment impact analysis (NSIA) framework. International Workshop on Enterprise Applications and Services in the Finance Industry,2016
[10] Wan, X., Yang, J., Marinov, S. *et al.* Sentiment correlation in financial news networks and associated market movements. *Sci Rep* 11, 3062 (2021).
[11] Bharathi.Sv, Shri & Geetha, Angelina. (2017). Sentiment Analysis for Effective Stock Market Prediction. International Journal of Intelligent Engineering and Systems.
[12]Anita Yadava, C K Jhaa , Aditi Sharanb , Vikrant Vaish, "Sentiment analysis of financial news using unsupervised approach", International Conference on Computational Intelligence and Data Science (ICCIDS),2019
[13] P. Uhr, J. Zenkert and M. Fathi, "Sentiment analysis in financial markets: A framework to utilize the human ability of word association for analyzing stock market news reports," 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2014.
[14] Turney, Peter D. (2002) "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania: 417-424.
[15] Michele Costola,Michael Nofer,Oliver Hinz,Loriana Pelizzon,,"Machine learning sentiment analysis, COVID-19 news and stock market reactions" ,
2020
[16] D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018
[17] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208.
[18] László Nemes & Attila Kiss (2021) Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication
[19] Joshi, Kalyani & N, Bharathi & Rao, Jyothi. "Stock Trend Prediction Using News Sentiment Analysis". International Journal of Computer Science and Information Technology. 2016, pages-. 67-76
[20] Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan, Maitreyee Dutta,
Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty,Procedia Computer Science,Volume 70,2015,Pages 85-91.