Correlation Analysis - Stock Market & Social Media Trends

Group ID: G5_8

Nuzair Mohamed S
Sanyog Kave P

Dr. N. Sujaudeen
Associate Professor

SSN College of Engineering, Chennai

February 10, 2023

Outline

- **Motivation**
- **Problem Statement**
- **Justification**
- **Feasibility study**
- **Architectural Design**
- **Expected Outcomes**
- **Project Timeline Chart**
- **Literature Survey**
- **References**

Motivation

- In recent years, it is seen that stock values are highly influenced and impacted by news events.
- Social media trends are increasingly influencing the stock valuations as seen from time to time.
- Due to this, we see frequent and unexpected stock market fluctuations more recently.
- Immediate reactions to the stock market can be seen when incidents are reported or propagated using extreme sentiments.

Problem Statement

- In a nutshell, we wish to analyse the correlation between social media trends & news articles relating to specific sectors and its effect on their current stock valuations.
- To develop a big-data architecture that is capable of handling real-time streaming information from news sources like Google News, Economic Times, Yahoo Finance, as well as renowned social media platforms like Twitter, Reddit etc.
- To assess and quantify the effect of news and social media on the economics of different sectors and hence its influence on the stock market.

Justification

- A huge margin of people invest in the stock market as a means to make money quickly and their reactions towards their owned stock of relevant companies based on news trends are adverse and almost immediate, making their valuations more volatile.
- A lot of young people have started to invest and are very influenced by social media trends.
- Presently, in this fast paced digital world being online, people tend to be more engaged on social media & use it as a source of credible information. Finding the correlation between such events & stock valuations is hence justified.

Feasibility study

- **Technical Feasibility**

  1) The feasibility of the project is based on the research done by Lee et al., where they collect Tweets and news from specific sectors to find a correlation between the expressed Tweet sentiments and the stock value.

  2) Our seniors (Shashanka, Venkataraman, Vishakan) have built a similar model and tried to correlate tweets' impact on stock market.

  These are proof for the feasibility and hence motivates us to build a better model with inclusion of various other factors like news articles, reddit, etc.

- **Financial Feasibility**

  This project requires almost no financial support as most of data sources and tools are open-source and are easily acquirable.

Architectural Design

- **Proposed System**

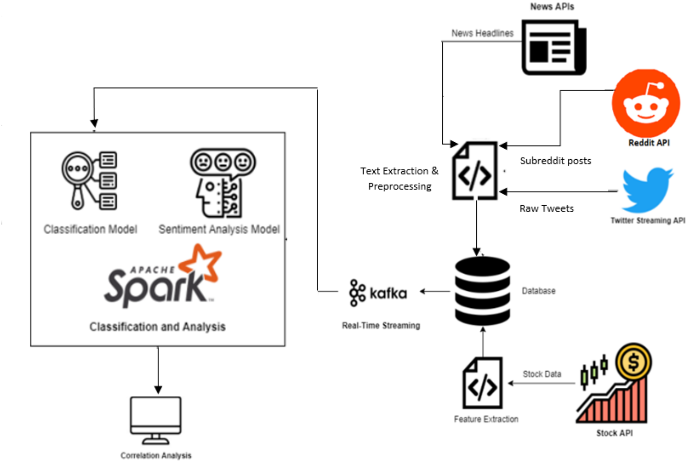


Figure: Proposed architecture of the system

- **Category Classifier Model**

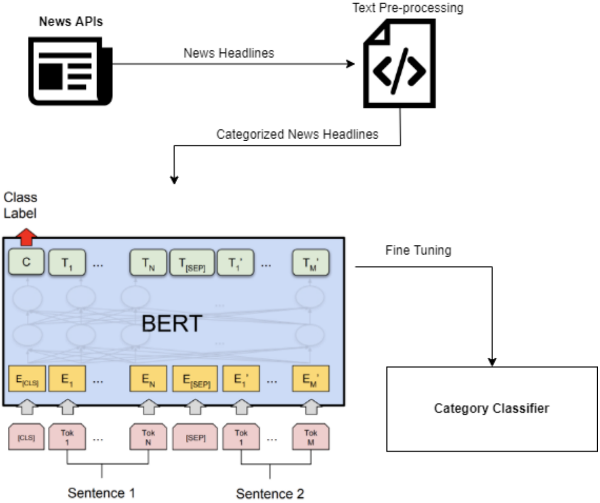


Figure: Fine tuning the BERT model using News data

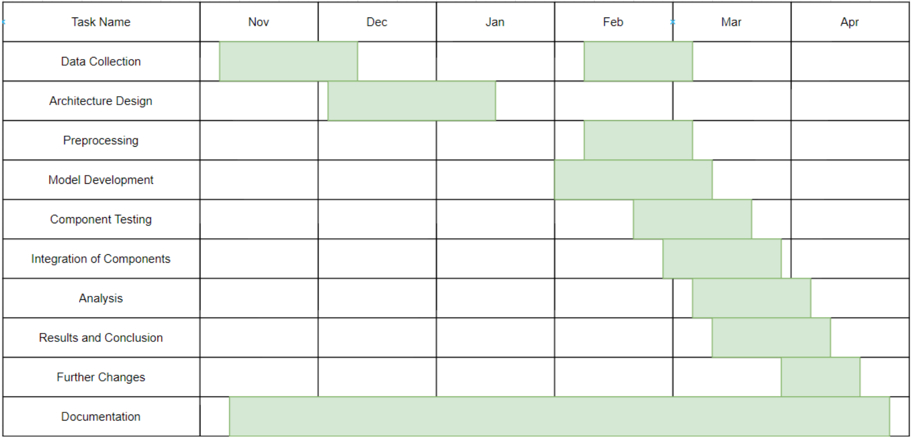
- **Data Collection and Processing**
  - **Twitter Data Processing**
    The Twitter Streaming API (GET https://api.twitter.com/2/tweets/search/recent) along with keywords for each category of analysis was used to collect tweets over the span of January and February of 2023. A Python script was written to collect the data, and generate a dataset of tweets labelled according to their category.
  - **News Data Processing**
    Three Python scripts were used to collect data from each of the APIs shown below and collect the combined data in a CSV file. The Headlines were labelled based on the the category they were extracted as. 1. NewsData: GET https://newsdata.io/api/1
    2. NewsCatcher: GET https://api.newscatcherapi.com/v2/search
    3. NewsApi: GET https://newsapi.org/v2/everything?
  - **Reddit Data Processing**
    The Reddit API (GET https://www.reddit.com/api/v1/access_token) was used to collect data from the social media platform Reddit. A python script was used to extract data from the subreddits and saved in a CSV file according to their extracted category.

- **Data Collection and Processing**
  - **Stock Data Processing**
    The Polygon API (GET https://api.polygon.io/v1/open-close) was used to collect stock ticker data. The API accepted the type of ticker data required, and the date to get the data from. A Python script was used to collect Stock data for the span of January and February of 2023, and was stored in a dataset categorized according to the sector of the companies queried.
  - **Storing in database**
    Since all data were extracted in a neatly categorized format, they can now be stored in a relational database like SQLite to facilitate further processing by the real-time streaming tools.

Expected Outcomes

The main purpose of the entire analysis, is to unearth a correlation between information from social media and news articles and stock market changes, if any. The expected outcome, therefore, is an accurate correlation analysis betweeen social media trends, and the stock market variations in the related sectors and to infer if social media influences stock market.

Project Timeline Chart



Task Name	Nov	Dec	Jan	Feb	Mar	Apr
Data Collection	🟩	🟩		🟩	🟩	
Architecture Design		🟩	🟩			
Preprocessing				🟩	🟩	
Model Development				🟩	🟩	
Component Testing				🟩	🟩	
Integration of Components				🟩	🟩	
Analysis					🟩	🟩
Results and Conclusion					🟩	🟩
Further Changes					🟩	🟩
Documentation	🟩	🟩	🟩	🟩	🟩	🟩

Figure: Timeline Chart

Literature Survey

1) Lee et al. : It delves into analyzing Twitter data, by classifying it into categories and performing sentiment analysis to predict the trend of stock prices, and comparing it to reality, to find the correlation between the two.
2) Shashanka et al. : Increase in volume of latest data from twitter and improved ML model.

Scope of improvement :

- News articles and platforms like Reddit also affect stock market.
- New huge sets of data from different sources as well as updated data-set from twitter.
- Integration of data from multiple sources require better ML model to improve classification and sentimental analysis.

References

[1] Chungho Lee, Incheon Paik, Stock Market Analysis from Twitter and News Based on Streaming Big Data Infrastructure

[2] Social media based Stock Market analysis using big-data infrastructure by Shashanka Venkatesh, Venkataraman Nagarajan, Vishakan Subramanian

[3] Vaanchitha Kalyanaraman, Sarah Kazi, Rohan Tondulkar, Sangeeta Oswal, Sentiment Analysis on News Articles for Stocks

[4] Zhihao Peng, Stocks Analysis and Prediction Using Big Data Analytics, 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS).

[5] Kanavos, Andreas & Vonitsanos, Gerasimos & Mohasseb, Alaa & Mylonas, Phivos. (2020). An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space

[7] Marsland, Stephen, Machine Learning: An Algorithmic Perspective, 2014

[8] Apache spark and kafka documentation

References

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[11] Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen, BERTweet: A pre-trained language model for English Tweets

THANK YOU