**Citations**

1. Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.
2. Chen, R., & Lazer, M. (2013). Sentiment analysis of twitter feeds for the prediction of stock market movement. *stanford edu Retrieved January*, *25*, 2013.
3. Patel, A. P., Patel, A. V., Butani, S. G., & Sawant, P. B. (2017). Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches. *IJIRST-International journal for Innovative Reasearch in Scinece & Technology*, *3*(10).
4. Varghese, R., & Jayasree, M. (2013). A survey on sentiment analysis and opinion mining. *International journal of Research in engineering and technology*, *2*(11), 312-317.
5. Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2), 1-41.
6. <https://www.dailyfx.com/education/understanding-the-stock-market/stock-market-sentiment-analysis.html>
7. Sun, Y., Wu, M., Zeng, X., & Peng, Z. (2021). The impact of COVID-19 on the Chinese stock market: Sentimental or substantial? *Finance Research Letters*, *38*, 101838.
8. Yoon, B., Jeong, Y., & Kim, S. (2020). Detecting a Risk Signal in Stock Investment Through Opinion Mining and Graph-Based Semi-Supervised Learning. *IEEE Access*, *8*, 161943-161957.
9. Paramanik, R. N., & Singhal, V. (2020). Sentiment Analysis of Indian Stock Market Volatility. *Procedia Computer Science*, *176*, 330-338.
10. Al-Awadhi, A. M., Alsaifi, K., Al-Awadhi, A., & Alhammadi, S. (2020). Death and contagious infectious diseases: Impact of the COVID-19 virus on stock market returns. Journal of behavioral and experimental finance, 27, 100326.
11. Ashraf, B. N. (2020). Stock markets’ reaction to COVID-19: Cases or fatalities?. *Research in International Business and Finance*, *54*, 101249.
12. Shan, L., & Gong, S. X. (2012). Investor sentiment and stock returns: Wenchuan Earthquake. *Finance Research Letters*, *9*(1), 36-47
13. Baig, A. S., Butt, H. A., Haroon, O., & Rizvi, S. A. R. (2021). Deaths, panic, lockdowns and US equity markets: The case of COVID-19 pandemic. *Finance research letters*, *38*, 101701.
14. Ichev, R., & Marinč, M. (2018). Stock prices and geographic proximity of information: Evidence from the Ebola outbreak. International Review of Financial Analysis, 56, 153-166.
15. <https://www.gtarsc.com/>
16. <https://ifind.bnu.edu.cn/>
17. Chapelle, O., Schölkopf, B., & Zien, A. (2006). A discussion of semi-supervised learning and transduction. In *Semi-supervised learning* (pp. 473-478). MIT Press.
18. Hussain, A., & Cambria, E. (2018). Semi-supervised learning for big social data analysis. Neurocomputing, 275, 1662-1673.
19. Yu, D., Chen, N., Jiang, F., Fu, B., & Qin, A. (2017). Constrained NMF-based semi-supervised learning for social media spammer detection. Knowledge-Based Systems, 125, 64-73.
20. Rout, J. K., Dalmia, A., Choo, K. K. R., Bakshi, S., & Jena, S. K. (2017). Revisiting semi-supervised learning for online deceptive review detection. IEEE access, 5, 1319-1327.
21. Lašek, I., & Vojtáš, P. (2011). Semantic information filtering-beyond collaborative filtering. In 4th International Semantic Search Workshop. Retrieved from: http://km. aifb. kit. edu/ws/semsearch11/11. pdf.
22. Moseki, K. K., & Madhava Rao, K. S. (2017). Analysing stock market data—Market sentiment approach and its measures. Cogent Economics & Finance, 5(1), 1367147.
23. Rao, K. M., & Ramachandran, A. (2014). Exchange rate market sentiment analysis of major global currencies. *Open Journal of Statistics*, *2014*.
24. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.
25. Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, 4(2), 120-137.
26. Kuzmina, J. (2010). Emotion's component of expectations in financial decision making. Baltic Journal of Management.

**Abstract**

In this paper we have tackled the forefront of sentimental analysis used in the stock markets. Sentimental analysis refers to the contextual understanding of the desired target, this can be done by using machine learning approach which will classify and analyse the target’s sentiments, behaviour, opinions and emotions, etc, about a specific topic (here, the stock market) which they have conveyed via their speech, text, etc. This understanding of text is done with the help of NLP. The stock market is a place where people take part in a continuous exchange (buying and selling) of shares. It has always been a place to earn money and during the pandemic it has experienced a great boost in the number of people interacting with market. In this survey paper, we have analysed and given our understanding of the topic; which includes the study on a large collection of research papers, articles and existing systems on the topic. Sentimental analysis can be used in stocks in hopes to determine an investor’s opinion on a certain group/set of assets or stocks. By understanding these opinions, the analyst may be able to predict the future price action. If a person is able to understand the opinions and sentiments of the general public, he/she would be able to make a prediction of high accuracy. This literature survey is done to study the sentiment analysis of stock market in-depth and to familiarize with other works done on the subject

**Keywords:**

Sentimental analysis, stock, stock market, prediction, investor sentiment, opinions , machine learning, NLP, shares

**Your Contribution:**

problem analysis (as discussed by the authors in the chosen paper), method proposed, and evaluation of the proposed method)

Detailed description of topic, Proposed Solution -reason, Architecture, approach, results

In our expedition of understanding the ‘sentimental analysis in stock prediction’ we have surveyed a number of papers. In the paper, ‘**The Impact of COVID-19 on the Chinese Stock Market: Sentimental or Substantial?**’ by the authors Sun Y., Wu M., Zeng. X., and Peng Z., they have investigated the impact on the Chinese Stock Market caused by COVID-19 by doing an event study and examine the effects of individual investor sentiment on their returns. This paper serves as a good basis to understand the sentimental analysis involved in stocks because during the pandemic not only was there a huge rollercoaster of stock prices but also a significant increase in the number of traders / investors in the market.

During The COVID-19 pandemic, which began in early 2020, has caused financial market turmoil. Circuit breakers struck the American stock market twice in a week, and the situation in other countries was not much better. The majority of academics have noticed drops throughout the epidemic, but the causes are unknown. It is suspected that region with a higher number of confirmed cases would suffer more significant losses. Naturally, that sector's profitability would be harmed, and its stock returns would suffer as a result, but this hypothesis was proven wrong. Their research demonstrates that this is not the case. The stock returns of Hubei businesses are identical to those of the market. Pharmaceutical stocks' high returns did not endure as long. This oddity supports the theory that stock market volatility during the COVID-19 outbreak was not solely due to economic loss.

In this paper it was found that, Individual investor emotion is positively connected with stock market returns during the outbreak, according to the findings. Stocks with high PB, PE, CMV, net asset, and institutional shareholder ratios, as well as extended listed years, are more susceptible to the epidemic. In consideration of the above stated reasons, they have used a Panel Regression Model. This model is a two-dimensional construct in which the same participants are observed multiple times across different time periods. Panel data is a blend of cross-sectional and time-series data in general. One observation of many objects and accompanying variables at a single point in time (i.e., an observation is taken once) is described as cross-sectional data. In their approach they have also made use of the Fama-French model. This model attempts to explain stock returns using three factors: (1) market risk, (2) small-cap outperformance against large-cap outperformance, and (3) high book-to-market value outperformance versus low book-to-market value outperformance. The advantages of the proposed work are that they have used panel regression implying that they have tested their data over a period of time to give a proper solution which is helpful for performing sentimental analysis.

In their approach the authors used two particular datasets, (1) was a stock-related financial data are from the CSMAR database and another (2) was a sentiment data used in this work is GubaSenti, on which they applied three main equations to help find their result. (a) , this equation gives the ordinary least squares (OLS) regression where, Ri,t represents the return of index i on date t in the estimation window, and MKTi, SMBt and HMLt are the three factors of the Fama–French model.   
(b)  , this equation gives abnormal returns where, Rt represents the actual return on date t in the event window. (c)  , this equation gives individual abnormal returns to create a “cumulative abnormal return (CAR)”.

After experimenting with their approach, they came up with two results (1) The cumulative abnormal return in the event window is positive, indicating that the outbreak has a strong short-term positive impact on the stock price. The second result, regarding pharmaceutical stocks, reveals that the t-value is significantly positive, showing that the epidemic has a strong beneficial impact on pharmaceutical stock prices. And (2) that the findings show that mood can have a big impact on the overall market return during an epidemic. It also backs up the idea that the reverse effect is strong, implying that stock returns fell during the post-event period.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The next paper we analyzed was ‘**Detecting a Risk Signal in Stock Investment Through Opinion Mining and Graph-Based Semi-Supervised Learning’** by the authors Yoon, B., Jeong, Y., & Kim, S. they have turned their head towards the problem as to how one can avoid credit events that might cause a national and global economic crisis which will ultimately lead to socio - economic losses. In this paper, they have shown how the majority of global economic crises are generated by a cascade of tiny events that have a potentially large impact. We can prevent a seriously damaging national or global crisis by recognizing and catching these tiny occurrences before they happen if we can notice and catch them before they happen. There has also been a surge in a number of hazy phenomena, such as cryptocurrency, a new financial service. The widespread practice of taxing imports to protect a country's native industries from outside competition leads to trade wars between countries and, in the long run, may hamper economic progress. All these factors can contribute to a credit event which is defined as an incident that seriously affects the bankruptcy risk of a company. Therefore, it is necessary to pre-determine the factors affecting the credit score so we can work towards monitoring and prevention of the same.

The study's main goal is to use opinion mining and graph-based semi-supervised learning to create an algorithm to aid in stock investment decision-making. Because of the massive increase in data in recent years, not only analysts and professionals, but also individual investors, may now acquire superior-quality financial and non-financial data about companies. This data can be an important source for detecting market moves. As a result, the goal of this work was to use opinion mining and machine learning to construct an algorithm to aid in stock investment decision-making using both objective and subjective information.

The three sections of the algorithm established in this study were as follows: (1) data gathering and filtering, (2) credit risk assessment and early warning signal identification, and (3) credit event prediction First, data was gathered from a variety of stock-related databases, ranging from news and financial statements to social networking sites and online communities. Author analysis and a rule-based method were used to filter bogus material, such as rumors and fake news. Second, sentiment analysis and opinion mining recognized a risk signal, which is an indicator or trigger of credit events such as bankruptcy and delisting. The risk signal was defined by three grades (dangerous', warning', and `caution') in stock investment to provide insights for monitoring and responding to credit events in advance. Third, the likelihood of credit events happening was predicted using logistic regression, which included a binary dependent variable (occurring or not occurring) and independent factors based on signal detection findings.

The authors of this paper have proposed, a novel algorithm to recognize risk signals and anticipate the future occurrence of credit events to aid in stock investing decision-making. Sentiment analysis based on opinion data, word2vec, and graph-based semi-supervised learning may be used to detect the danger signal by evaluating the sentiment value of data, including news and views. A logistic regression model comprised of indicators based on the sentiment value of views then predicts the likelihood of credit occurrences. To create a logistic regression model for predicting future events, data from companies in the same industry, such as the sentiment value of views collected from surveys, were used. To do so they have performed sentimental analysis based on opinionated data which includes logistic regression model, linguistic rule- based model and a graph-based semi-supervised learning.

They have also utilized four main equations (a) , this equation provides us with the sentimental value of the keyword. (b) [ Pi,j = \*(Sentimental value of core keyword i) ], which gives the proximity index between keyword i and core keyword j, (c) [ = (j=1,2,…m) ], is an equation for the Euclidean distance between keyword i and core keyword j and (d) I= ,in which ‘I’ stands for Indicator for monitoring.

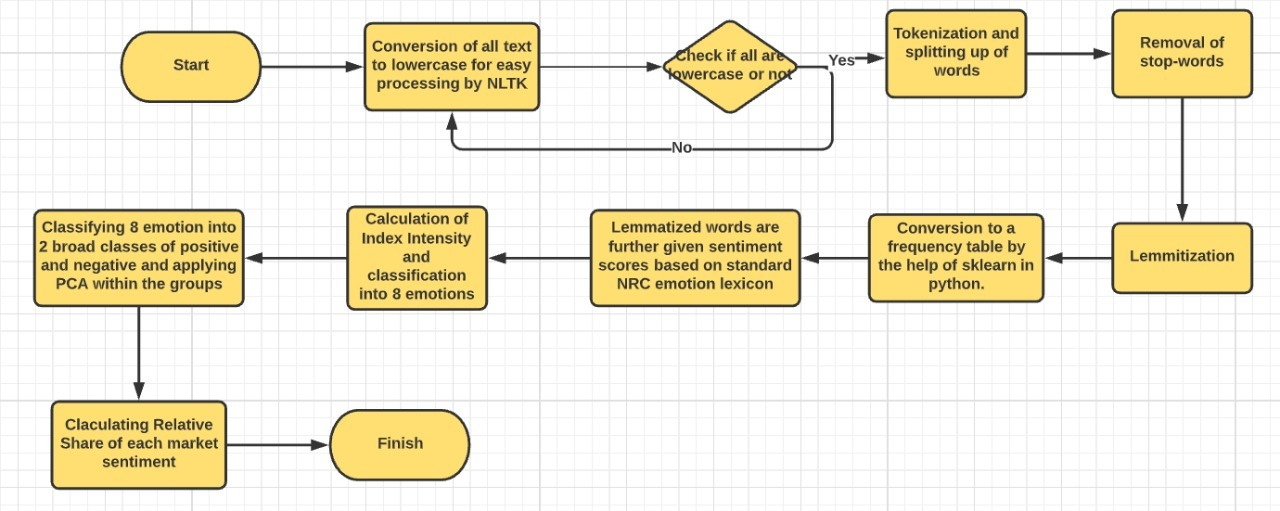
This paper suggests behaviour- and language-based approaches in sequence, which means that different characteristics of language in fake or genuine opinions are inspected by sentiment analysis after examining the current trend of opinion occurrences. The behaviour-based approach aims to identify the distribution of opinions by investors, while the language-based approach can pinpoint the pattern of opinions. To estimate the sentiment value of words, the document is preferentially rated, which is conducted in two ways: stock price and review score. The rating differs depending on whether the company is listed on a securities market or not. After deciding the sentiment of each article, this sentiment value is disseminated to all words included in the article through Naïve-Bayes classification, which is based on the co-occurrence of words in the article. Using the logistic regression equation, data of the target firm are put into the prediction model based on logistic regression. The probability of credit event occurrence is finally estimated. To validate the results derived from this prediction model, the authors developed a confusion matrix by comparing the actual number of incidences with the predicted number, which is higher than the cut-off probability and represented as 0 or 1 (binary).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In the paper, ‘**Sentiment Analysis of Indian Stock Market Volatility**’, by the authors Paramanik, R. N., & Singhal, V., have presented the issue of scarcity of literature in the context of India’s stock market volatility using investor’s sentiment analysis. This paper attempts to shape the volatility of Indian Stock market using investor sentiment analysis. The recent literature in the behavioral finance has challenged the notion of a rational investor in the market since the emergence of noise traders in the market due to their cognitive errors and emotional exuberance. The transitory influence of noise traders were first believed to be eradicated but the argument of traditional financial theory is challenged by many researchers Normally, sentiment is understood as the overall attitude of an investor’s behavior but the influence of such market volatility has been fixed in literature by proxies of market analysis. There is scarcity of literature in the context of Indian stock market’s volatility using sentiment analysis.

Three different generalized autoregressive conditional heteroscedasticity (GARCH) models are used to analyze impact of market sentiments. Emotional and sentiment indices are constructed by the help of NLP techniques. Initially the data is gathered and the headlines and summaries are gathered from the resources and NLTK (Natural Language Toolkit) is used to filter the data. In the first step we covert all the texts to lower case so that NLTK can process it easily. Next tokenization is done so that the words in the string. Next, Stop-words are removed from the string by the help of NLTK so that common words can be removed that do not generate any significance. In the further step, Lemmatization is done on various similar split words to get one common word as this helps us get better frequency insights. The pre-processed data is then converted to a frequency table using sklearn feature extractor of python. This table contains date wise frequency of each lemmatized word. These lemmatized words are further given sentiment scores based on standard NRC emotion lexicon (Emolex) consisting of 14,181 words with eight basic emotions (i.e., anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Now the mentioned 8 emotions are further classified into two broad categories which are positive and negative sentiments. Anger, fear, sadness and disgust are negative sentiments and others are positive sentiments. Principal component analysis (PCA) is employed to respective groups of emotions where derived factor loading is assigned as suitable weights for each emotion to construct scores for the two sentiments. At last, the relative share of each type of sentiment is calculated and relative dominance of two types of market sentiments is measured.

The approach used in this paper is known as NLTK with volatility modelling. The authors use the algorithm of Natural Language Toolkit (NLTK) and Principal Component Analysis (PCA) in order to complete the proposed work. Sklearn feature extractor is also used in order to complete the proposed work. The architecture of their work can be seen below



This paper attempts to augment the existing GJR GARCH model with previously proposed variables, share of positive market sentiment POS) and share of negative market sentiment (NEG) to assess more rigorously how these two contradictory sentiments shape the dynamics of conditional volatility using the augmented conditional volatility equation as shown in the equation section. When we analyse the finding it is very much clearly evident that the noise traders play a dominant role. An advantage of finding is use of real stream of data which is more useful in daily life. This stream of data is found to be more dynamic in nature than traditional monthly and quarterly indicators. Also, the given model does not consider positive and negative errors in mean equation and instead of that the model has generated separate market sentiments on conditional volatility of Indian financial market. This approach is better and more appealing in today’s financial market.