STOCK MARKET ANALYSIS

Anoush Tittu T, Rakshitha K, Nanditha TN, Sireya Rani M, Yukthi S R

Department of Computer Science and Engineering

Presidency University, Bangalore, India

ANOUSH.20201CSE0658@presidencyuniversity.in, RAKSHITHA.20201CSE0706@presidencyuniversity.in

SIREYA.20201cse0707@presidencyuniversity.in , NANDITHA.20201CSE0708@presidencyuniversity.in

YUKTHI.20201CSE0678@presidencyuniversity.in

***Abstract*--** The abstract aims to address the correlation between stock market movements and public sentiments expressed on Twitter. It delves into the utilization of sentiment analysis and supervised machine learning techniques to explore this connection. The study leverages Word2vec for textual representation, examining how shifts in stock prices align with sentiments expressed in tweets about specific companies. The investigation underscores the potential impact of positive news and social media sentiments on stock prices, emphasizing a demonstrated correlation between fluctuations in stock prices and sentiments conveyed in Twitter.-

Keywords:

Hashtag Collection, Data Collection, Real-Time Stock History Data, Positive Keywords, Negative Keywords, Polarity Computation, Sentiment Analysis , Sentiment Index Computation, Sentiment Discrepancy Index, Price Prediction, Yahoo Finance API

I. Introduction

The stock price, a form of time series data within the financial domain, exhibits dynamic, selective, and nonlinear fluctuations, making accurate forecasting challenging. The utilization of data mining or machine learning techniques to predict stock prices has emerged as a significant concern in recent times. However, predicting stock prices based on the efficient market hypothesis is hindered by the random walk pattern they tend to follow.

A stationary prediction strategy is also impractical, as successful forecasting rules would likely lead to self-destruction once discovered by investors. Various factors contribute to the fluctuations in financial market movement, rendering predictions of stock market prices and directions difficult. Statistical analysis methods show promise in identifying key factors impacting short-term stock volatility, while data mining techniques have proven effective in generating highly accurate predictions of stock price movements.

Many financial analysts and stock market investors believe in the potential profitability of employing technical analysis approaches for stock market forecasting. However, historical stock prices, once the basis for stock market prediction, have been debunked in later studies due to the inherent fluctuation in stock market prices. The efficient market hypothesis asserts that financial market movements are influenced by news, current events, and product releases, making predictions challenging.

Given the unpredictable nature of news and current events, stock market prices are considered to follow a random walk pattern and cannot be predicted with more than 60-65% accuracy. The advent of social media has introduced a wealth of information about public sentiments. Platforms like Twitter, with over 140 million tweets daily, have become valuable for researchers, offering concise public opinions on various topics. The information extracted from tweets has proven useful for making predictions.

II. LITERATURE SURVEY

In the paper titled "Efficient Market Hypothesis and Forecasting," the author argues that the Efficient Market Hypothesis (EMH) introduces forecasting tests that resemble those applied when assessing the optimality of a forecast within a specified information set. However, notable distinctions arise due to market efficiency tests relying on identifying profitable trading opportunities in 'real time.' Forecasters are continually seeking predictable patterns, impacting prices as they endeavor to exploit trading opportunities. The transitory nature of stable forecasting patterns becomes apparent, as they are unlikely to persist over extended periods, leading to self-destruction upon discovery by a large number of investors. This emergence of nonstationarities in the time series of financial returns complicates both formal tests of market efficiency and the pursuit of successful forecasting approaches.

In the paper entitled "The Efficient Market Hypothesis and Its Critics," the author posits that revolutions often give rise to counterrevolutions, and the efficient market hypothesis in finance is no exception. The once dominant intellectual paradigm of the efficient-market revolution has faced challenges, particularly from economists emphasizing psychological and behavioral factors in stock-price determination. Additionally, econometricians argue that stock returns are, to a significant extent, predictable .This survey delves into the critiques directed at the efficient market hypothesis and explores the intricate relationship between predictability and market efficiency. The author concludes that, contrary to some recent academic papers, our stock markets exhibit greater efficiency and less predictability than suggested.

In the paper titled "Forecasting with Artificial Neural Networks: The State of the Art [j]," the author contends that the growing interest in employing artificial neural networks (ANNs) for forecasting has sparked a significant increase in research activities over the past decade. While ANNs hold considerable promise, they also encompass a notable degree of uncertainty. Researchers, as of now, remain uncertain about the impact of key factors on the forecasting performance of ANNs. This paper aims to present a state-of-the-art survey of ANN applications in forecasting. The objective is to offer a synthesis of published research in this domain, insights into ANN modeling issues, and indications of future research directions.

In the paper titled "Community Detection and Mining in Social Media [J]," the author aims to highlight the transformative impact of participatory web and social media over the past decade. This period has witnessed the convergence of millions of users engaging in various activities online, such as playing, tagging, working, and socializing, leading to novel forms of collaboration, communication, and intelligence that were previously unimaginable. Social media has not only reshaped business models but has also influenced opinions, emotions, and provided unprecedented opportunities to study human interaction and collective behavior on a massive scale.
Approaching the subject from a data mining perspective, the paper introduces the characteristics of social media, examines representative computing tasks associated with social media, and addresses the challenges inherent in these endeavors. The lecture covers basic concepts, presents state-of-the-art algorithms with easily understandable examples, and advocates for effective evaluation methods. The focus is particularly on graph-based community detection techniques and their important extensions, which deal with dynamic and heterogeneous networks in social media. The paper also showcases how the identified community patterns can be leveraged for social media mining.

The concepts, algorithms, and methods presented in this paper serve as a guide to harnessing the potential of social media and developing socially-intelligent systems. This accessible introduction to the study of community detection and mining in social media is intended for students, researchers, and practitioners in various disciplines and applications where social media plays a crucial role as a key data source, fueling curiosity and promoting understanding, management, innovation, and excellence.

In the paper titled "Text and Structural Data Mining of Influenza Mentions in Web and Social Media [J]," the author aims to demonstrate the value of text and structural data mining in web and social media (WSM) for disease surveillance. This approach not only serves as a unique resource but also facilitates the identification of online communities, enabling targeted public health communications (PHC) for the widespread dissemination of relevant information. The paper illustrates how text mining can identify trends in flu-related posts that correlate with real-world influenza-like illness patient reports. Additionally, a graph-based data mining technique is introduced to detect anomalies within flu-related blogs based on factors such as publisher type, links, and user-tags.

In the paper titled "Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards; Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)," the author aims to develop a methodology for extracting sentiment from stock message boards, specifically focusing on small investor sentiment.

The study employs five distinct classifier algorithms in conjunction with a voting scheme, demonstrating effective performance compared to human and statistical benchmarks. The quality of the resultant sentiment index is enhanced through time series and cross-sectional aggregation of message information.

Empirical applications of the methodology reveal a relationship between sentiment and stock returns, assessed through visual analysis, phase-lag analysis, pattern recognition, and statistical methods. The paper suggests that sentiment possesses an idiosyncratic component, and aggregating sentiment across stocks aligns more closely with index returns than individual stock sentiments. The preliminary evidence also indicates that market activity can influence small investor sentiment. Consequently, the algorithms developed in this paper offer potential applications for assessing the impact of management announcements, press releases, third-party news, and regulatory changes on investor opinions. The author acknowledges the supportive environments at UC Berkeley's Computer Science Division and Haas School, where this work was initiated, and expresses gratitude to David Levine for valuable comments and providing the title.

In the paper titled "Closed-end Country Funds and US Market Sentiment [J], Review of Financial Studies," the author aims to explore the phenomenon where closed-end country funds often trade at significant premiums or discounts from their foreign asset values (NAVs). The investigation into this anomaly reveals that individual fund premiums exhibit simultaneous movements, primarily influenced by the co-movement of their stock prices with the U.S. market.

Additionally, the paper introduces an index of country fund premiums, which proves effective in differentiating size-ranked U.S. portfolio returns and predicting country fund stock returns. These findings suggest that the prices of international equities are impacted by local risk. Specifically, the study demonstrates that movements in country fund premiums reflect a U.S.-specific risk, implying a connection with U.S. market sentiment.

In the paper titled "HHMM-based Chinese Lexical Analyzer ICTCLAS; Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume," the authors aim to present the results obtained from the Institute of Computing Technology, CAS, in the ACL SIGHAN-sponsored First International Chinese Word Segmentation Bake-off. The document introduces the unified Hidden Markov Model (HHMM)-based framework of their Chinese lexical analyzer, ICTCLAS, and provides an overview of the six tracks involved. The authors then present the evaluation results and offer additional analysis.

The evaluation on ICTCLAS indicates competitive performance, with the system ranking at the top in both CTB and PK closed tracks. In the PK open track, it secures the second position. Notably, the ICTCLAS BIG5 version was transformed from the GB version in just two days, yet it performed well in two BIG5 closed tracks. The authors reflect on the insights gained during the first bake-off, emphasizing the value of learning more about the development in Chinese word segmentation and gaining confidence in their HHMM-based approach. They acknowledge the challenges identified during the evaluation, expressing gratitude for the interesting and helpful experience provided by the bake-off.

In the paper titled "Semantic Orientation Computing based on HowNet," the author aims to address the growing significance of automated techniques for analyzing authors' attitudes towards specific events in the context of the evolving Internet and information explosion. These techniques are anticipated to play a pivotal role in business intelligence and public opinion surveys. Semantic orientation inference, as a meaningful tool, holds the potential to provide valuable information for various applications such as text classification, summarization, and filtering.

The paper focuses on measuring the semantic orientation of words, emphasizing its crucial role in predicting the author's attitude within a passage. The author introduces a straightforward HowNet-based method for computing the semantic orientation of Chinese words. Despite the method's reliance on only a few seed words, the paper asserts that satisfactory results can still be achieved. Notably, the performance of the method is reported to be particularly strong for frequently used words, with a frequency-weighted accuracy exceeding 80%.

In the paper titled "Text Opinion Mining to Analyze News for Stock Market Prediction [J]," the author aims to address the well-established connection between news and stock prices, highlighting the significant influence of news on stock market investments. Numerous research efforts have been dedicated to identifying this relationship or predicting stock market movements through news analysis. The recent utilization of massive news datasets, referred to as unstructured big data, has become increasingly popular for predicting stock prices.

The paper introduces a method focused on mining text opinions to analyze Korean language news, specifically for predicting rises and falls on the KOSPI (Korea Composite Stock Price Index). The methodology involves conducting Natural Language Processing (NLP) on news, outlining its features, categorizing and extracting sentiments, and identifying the opinions expressed by writers. The method then establishes correlations between news content and stock market fluctuations.

In the presented experiment, the paper demonstrates that the proposed method is effective in understanding unstructured big data. Furthermore, the research reveals that sentiments conveyed in news articles can be utilized to predict stock price fluctuations, irrespective of whether they are upward or downward. The extracted algorithmic experiments offer potential applications in making predictions about stock market movements.

III. Hardware & Software Requirement

Sl No	Parameter Name	Value
1	RAM	4GB
2	Hard Disk	360GB
3	Development Tool	Eclipse IDE
4	Front End Language/Technologies	Ext JS 4.2. & JSP
5	Server	Tomcat 7
6	Back End Language	JAVA,J2EE
7	Database Server	MySQL
8	Database Communication Framework	Spring JDBC
9	Front End and Backend Integration	Spring MVC

IV. Methodology

- ***Hashtag Collection:***
  Describes the real-time hashtags related to stocks that are collected from social media platforms.

- ***Data Collection using Tweets:***
  Outlines the process of collecting data from Twitter using hashtags.
  Based on the keywords, polarity will run against the tweet. So the tweets are collected through twexpertly (chrome extension)
  The collected data from the chrome extension is stored in a local database.
- ***Real-Time Stock History Data*** Sets from Yahoo Finance: Describes the use of Yahoo Finance API for collecting real-time stock-related information. Specifies the type of information that can be requested from the API.

- ***Positive Keywords and Negative Keywords***:
  Lists positive and negative keywords used in sentiment analysis.

- ***Polarity Computation Tweet Wise:***
  Explains the sentiment analysis process, including the use of positive and negative keywords.
  Defines how neutral polarity is computed.

- ***Sentiment Analysis Company Wise:***
  Associates each company with a set of tweets.
  Calculates positive, negative, and neutral polarity for each company based on its set of tweets.

- ***Sentiment Index Computation :***
  Describes the computation of the sentiment index for each tweet using a specified equation.
  $$SI = \ln \frac{1+N_{positive}}{1+N_{negtive}}$$

- ***Sentiment Discrepancy Index:***
  Explains the computation of the sentiment discrepancy index for stocks based on a given equation.
  $$SDI = \left|1 - \left|\frac{N_{positive}-N_{negtive}}{N_{positive}+N_{negtive}}\right|\right|$$

- ***Price Prediction of Stock:***
  Introduces the use of principal component analysis for predicting stock prices.
  Implies that factors influencing stock price trends are identified through this analysis.
  $$P_i = 0.6238 * P_{i-1} + 0.0455 * V_{i-1} + 0.0213 * TR_{i-1} + 0.0316 * M_{i-1} + 0.0423 * SDI + \beta$$

- ***Yahoo Finance API:***
  This library provides some methods that should make it easy to communicate with the Yahoo Finance API. It allows you to request detailed information, some statistics and historical quotes on stocks. Separate functionality is available to request a simple FX quote.

V. CONCLUSION

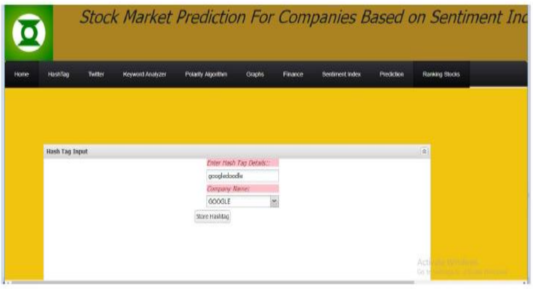
In this project, extensive work has been conducted on data mining algorithms, sentiment analysis,sentiment index creation, real-time stock prediction, and ranking of companies based on the stock market. A significant contribution of this work is the development of a sentiment analyzer capable of categorizing tweets into positive, negative, or neutral sentiments. The initial hypothesis posited that positive sentiments expressed by the public on Twitter about a company would correlate with its stock price. The obtained results strongly support this claim, suggesting a promising avenue for future research.

The innovative approach taken in this project provides users with a tool to predict the companies' order based on the propagation of their stock market increase. The integration of sentiment analysis into stock market predictions adds a valuable dimension to understanding market dynamics and investor sentiments. This work represents a meaningful contribution to the field, showcasing the potential for leveraging social media data for enhanced stock market insights and predictions.

VI. RESULT

Welcome Page

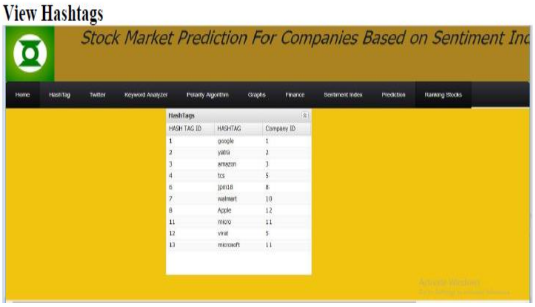
Hashtag Input



Hash Tag Input

Enter Hash Tag Details::
googledoodle
Company Name:
GOOGLE

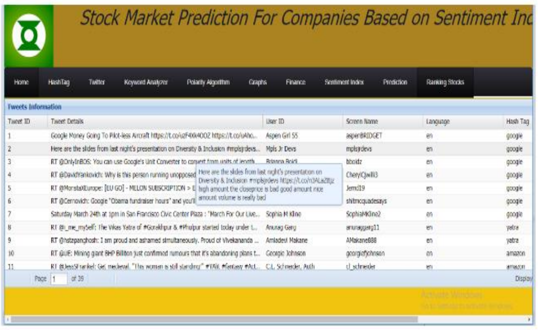
View Hashtags



HashTags

HASH TAG ID	HASHTAG	Company ID
1	google	1
2	yatra	2
3	amazon	3
4	tcs	5
6	jpm18	8
7	walmart	10
8	Apple	12
11	micro	11
12	virat	5
13	microsoft	11

View Tweets



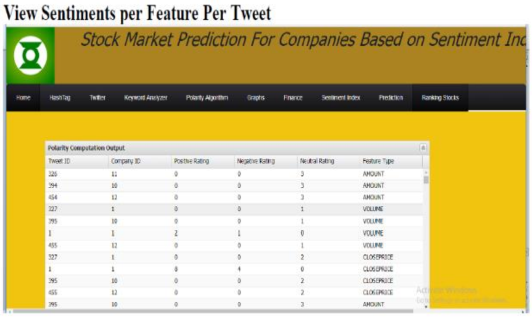
Tweets Information

Tweet ID	Tweet Details	User ID	Screen Name	Language	Hash Tag
1	Google Money Going To Pilot-less Aircraft https://t.co/uzF4Xk4OOZ https://t.co/uAho...	Aspen Girl 55	aspenBRIDGET	en	google
2	Here are the slides from last night's presentation on Diversity & Inclusion #mplsjrdevs...	Mpls Jr Devs	mplsjrdevs	en	google
3	RT @OnlyInBOS: You can use Google's Unit Converter to convert from units of length...	Brianna Boidi	bboidz	en	google
4	RT @DavidYankovich: Why is this person running unopposed		CherylCjwill3	en	google
5	RT @MorstaXEurope: [EU GO] - MELON SUBSCRIPTION > E		Jemd19	en	google
6	RT @Cernovich: Google "Obama fundraiser hours" and you'l		shitmcquadesays	en	google
7	Saturday March 24th at 1pm in San Francisco Civic Center Plaza : "March For Our Live...	Sophia M Kline	SophiaMKline2	en	google
8	RT @i_me_myself: The Vikas Yatra of #Gorakhpur & #Phulpur started today under t...	Anurag Garg	anuraggarg11	en	yatra
9	RT @hstzepanghosh: I am proud and ashamed simultaneously. Proud of Vivekananda ...	Amladevi Makane	AMakane888	en	yatra
10	RT @IJE: Mining giant BHP Billiton just confirmed rumours that it's abandoning plans t...	Georgie Johnson	georgiefjohnson	en	amazon
11	RT @JessSFrankel: Get medieval. "This woman is still standing" #YA #fantasy #Act...	C.L. Schneider, Auth	cl_schneider	en	amazon

Here are the slides from last night's presentation on Diversity & Inclusion #mplsjrdevs https://t.co/rOALaZBjz high amount the closeprice is bad good amount nice amount volume is really bad

Page 1 of 39 Display

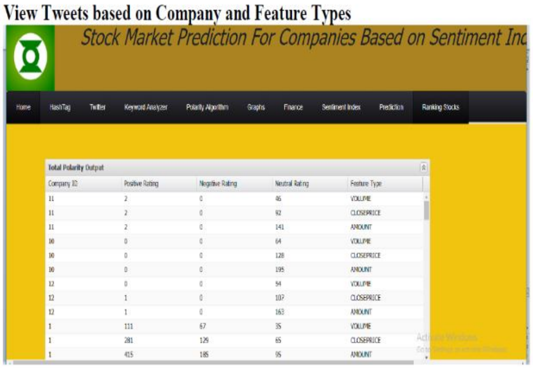
View Sentiments per Feature Per Tweet



Polarity Computation Output

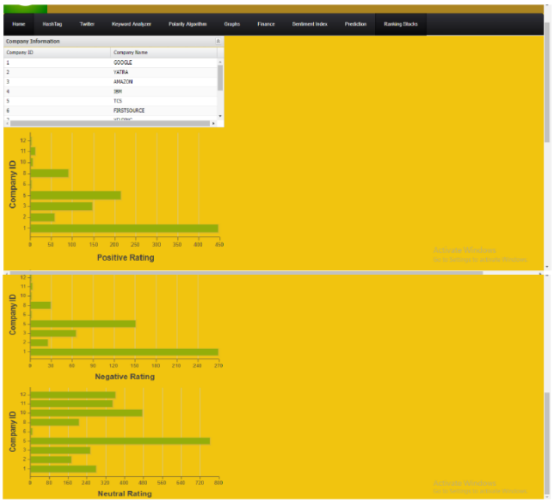
Tweet ID	Company ID	Positive Rating	Negative Rating	Neutral Rating	Feature Type
326	11	0	0	3	AMOUNT
394	10	0	0	3	AMOUNT
454	12	0	0	3	AMOUNT
327	1	0	0	1	VOLUME
395	10	0	0	1	VOLUME
1	1	2	1	0	VOLUME
455	12	0	0	1	VOLUME
327	1	0	0	2	CLOSEPRICE
1	1	8	4	0	CLOSEPRICE
395	10	0	0	2	CLOSEPRICE
455	12	0	0	2	CLOSEPRICE
395	10	0	0	3	AMOUNT

View Tweets based on Company and Feature Types



Total Polarity Output

Company ID	Positive Rating	Negative Rating	Neutral Rating	Feature Type
11	2	0	46	VOLUME
11	2	0	92	CLOSEPRICE
11	2	0	141	AMOUNT
10	0	0	64	VOLUME
10	0	0	128	CLOSEPRICE
10	0	0	195	AMOUNT
12	0	0	54	VOLUME
12	1	0	107	CLOSEPRICE
12	1	0	163	AMOUNT
1	111	67	35	VOLUME
1	281	129	65	CLOSEPRICE
1	415	185	95	AMOUNT



Company Information

Company ID	Company Name
1	GOOGLE
2	YATRA
3	AMAZON
4	IBM
5	TCS
6	FIRSTSOURCE

Sentiments Graphs

VII. FUTURE WORK

Expanding Data Sources:
Consider incorporating data from online reviews in addition to social media reviews. This broader data collection strategy can provide a more comprehensive understanding of public sentiment .

Algorithmic Improvements:
Introduce additional algorithms, especially focusing on data cleaning techniques, to further enhance the accuracy of sentiment analysis. Robust data cleaning processes can help in filtering noise and improving the overall quality of the dataset.

API Integration for Comparative Analysis:
Explore the use of the Google Finance API in conjunction with the Yahoo Finance API. This integration can facilitate a comparative analysis of stock data from different sources, potentially providing more comprehensive insights into market trends and stock performance.

VIII. REFERENCES

[1] TSAYY R S. Analysis of financial time series [M]. John Wiley & Sons,2005.

[2] TIMMERMANN A, GRANGER C W. Efficient market hypothesis and forecasting [J]. International Journal of Forecasting, 2004, 20(1):15-27.

[3] MAALKIEL B G. The efficient market hypothesis and its critics [J]. Journal of economic perspectives, 2003, 59-82.6

[4] ZHAANG G, PATUWO B E, HU M Y. Forecasting with artificial neural networks:: The state of the art [J]. International journal of forecasting, 1998, 14(1): 35-62.

[5] TAANG L, LIU H. Community detection and mining in social media[J]. Synthesis Lectures on Data Mining and Knowledge Discovery,2010, 2(1): 1-137.

[6] COORLEY C D, COOK D J, MIKLER A R, et al. Text and structural data mining of influenza mentions in web and social media [J].International journal of environmental research and public health,2010, 7(2): 596-615.

[7] DAS S, CHEN M. Yahoo! for Amazon: Extracting market sentiment from stock message boards; proceedings of the Proceedings of the Asia Pacific finance association annual conference (APFA), F, 2001 [C]. Bangkok, Thailand.

[8] BODURTHA J N, KIM D-S, LEE C M. Closed-end country funds and US market sentiment [J]. Review of Financial Studies, 1995, 8(3):879-918.

[9] EICHENGREEN B, MODY A: National Bureau of Economic Research, 1998.

[10] ZHANG H-P, YU H-K, XIONG D-Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS; proceedings of the Proceedings of thesecond SIGHAN workshop on Chinese language processing-Volume 17, F, 2003 [C]. Association for Computational Linguistics.

[11] ZHU Y-L, MIN J, ZHOU Y-Q, et al. Semantic orientation computing based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[12] KIM Y, JEONG S R, GHANI I. Text opinion mining to analyze news for stock market prediction [J]. Int J Advance Soft Comput Appl,2014, 6(1):

[13] SCHUMAKER R P, CHEN H. A discrete stock price prediction engine based on financial news [J]. Computer, 2010, 1): 51-6.

[14] ZHAOO L, WANG L. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm; proceedings of the Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on, F, 2015 [C]. IEEE