Rumor Project

Dursun Nurlu, Belgin Erdoğan, Hızır İlyas Aydoğan

Istanbul Technical University
Master of Science in Big Data & Business Analytics

Abstract

April 12, 2024

This paper introduces RUMOR, a platform tailored to monitor daily trends in publicly traded stocks. By integrating price fluctuations at specified intervals with data sourced from the Public Disclosure Platform (KAP) and harnessing social media engagement, RUMOR offers users comprehensive insights into market behavior. Additionally, RUMOR employs sentiment analysis by correlating stock-related discussions on social media with price movements, providing valuable indicators for investor sentiment. Moreover, for users lacking financial expertise, RUMOR provides an interpretive service for KAP news, offering nuanced perspectives on market events. This paper outlines the architecture, functionality, and potential applications of RUMOR, highlighting its significance in facilitating informed decision-making in the stock market.

Keywords: Stock market analysis, Social media analysis, Financial data processing, AI interpretation, Decision-making support, Data visualization, Investor dashboard

1. Introduction

The stock market has always played a central role in driving economies and fostering wealth creation. It has evolved into a dynamic arena where investors, and even curious individuals, seek opportunities. Even those who do not invest firsthand are often aware of the stock market's power, having heard news from their connections. With the advent of online trading platforms and easy access to information, participation in the stock market has dramatically increased. In today's world, where people check everything on their phones, stock market research has become easily accessible, sparking greater interest in this area. Consequently, many have incorporated stock trading into their daily routines, recognizing it as a dynamic system.

This surge in stock market participation owes much to technological advancements and the rise of innovative tools and platforms. Among these, RUMOR stands out as a game-changer, reshaping how investors analyze market trends. By tapping into real-time data from stock exchanges and blending it with insights from social media, RUMOR offers users a comprehensive view of market dynamics. This blend of traditional financial data with social media chatter gives investors a clearer picture, helping them make smarter decisions and spot new opportunities, as social media often reflects people's reactions to global events.

If we are taking into account Turkey stock market;"*The number of investors in the stock market has increased by 173.2% since the beginning of 2023, reaching 7.14 million as of September. The number of initial public offerings (IPOs) this*

year alone has reached 32. The money entering the stock market from these offerings has reached 42 billion TL."

This research aims to provide a solution for obtaining the most comprehensive information for the stock market for those who are interested. It is often the case that people struggle to simultaneously track stock prices, the latest KAP notifications, common trends, and individual thoughts. Through our data gathering methods and analysis, we facilitate the observation of multiple perspectives for investment. In today's technological age, we also incorporate AI into the equation, enabling it to provide commentary on KAP notifications in positive, negative, or mixed ways for the convenience of investors. Whether individuals seek a second opinion on their investments or simply lack expertise in this area, this additional insight can prove invaluable.

2. Literature Research

Everywhere you look, stockmarkets are breaking records. American equities, as measured by the sp 500 index, hit their first all-time high in more than two years in January, surged above 5,000 points in February and roared well above that level on February 22nd when Nvidia, a maker of hardware essential for artificial intelligence (ai), released spectacular results. The same day, Europe's stoxx 600 set its own record. Even before Nvidia's results had been announced, Japan's Nikkei 225 had surpassed its previous best, set in 1989. Little surprise, then, that a widely watched global stockmarket index recently hit an all-time high, too. [2]

Similarly, In 2024, according to data, the number of investors in Turkey's stock market reached 8.4 million, while the market capitalization soared to 13.2 trillion Turkish lira (TRY), setting a new record[4]

With the dramatically increasing number of participants in the stock market, it has become crucial for individuals to analyze stock data before making investment decisions. Various models applied for predicting the stock prices are managed using the time series models that involve Auto-Regressive Conditional Heteroscedastic (ARCH) model, Generalized Auto-Regressive Moving Average (GARCH), and Auto-Regressive Moving Average (ARMA). However, these models entail historical data and hypothesis like normality postulates. Several methods used for stock market prediction are based on conventional time series, such as fuzzy time series data, real numbers, and design of fuzzy sets. [3]

Web links are an important source of information. Also the Web is not only a huge repository of data and information but also a provider of services of all kinds. All these make the web a virtual society, where people, organizations and systems are interacting. Web mining is the process of discovering useful information or knowledge from hyperlink structure, pages content and data usage. There are three main Web mining tasks: web structure mining, web content mining and web usage mining. [1]

In this paper, the HTML parsing method is utilized as a data mining technique, which involves technically reading all data related to this study but parsing only the necessary ones. Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content, and translates it into a relational form, is called a wrapper. Wrapper generation algorithms assume that input pages of a wrapper induction system conform to a common template and that they can be easily identified in terms of a URL common scheme. [5]

3. Architecture

The steps taken are as follows: Data Collection, Data Storage and Processing, AI Analysis, and Data Visualization. The data architecture shown in Figure 1 will be discussed in the next section.

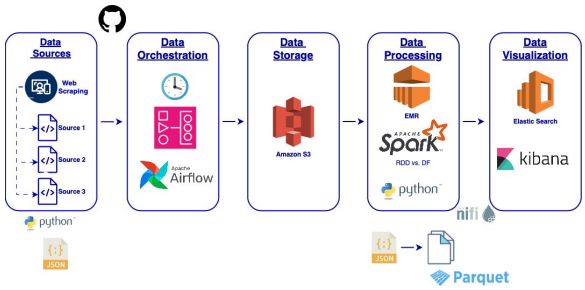


Figure 1. Architecture

3.1 Data Source

The data was collected from three main sources in a Python environment:

- BIST100 live price movements on an hourly basis, kept in CSV format. There is one CSV file for each day containing the BIST100 stocks and their hourly prices.
- KAP notifications on a daily basis for the BIST100 stocks of interest. Stock codes were used to match the notifications. This data handled in json format.
- Comments data gathered from social media forums on a daily basis, with stock codes used to match the comments. This data handled in json format.

3.2 Data Orchestration

In order to achieve streaming data, a workflow management platform is necessary. Given the necessity to obtain BIST100 stock prices data hourly and KAP notifications and comments daily, the "Airflow" platform was employed, and the environment was configured accordingly. Two Directed Acyclic Graphs (DAGs) were created to manage the distinct data scraping processes. By establishing this system, streaming data flow was successfully achieved and is now ready to be stored.

3.3 Data Storage

Following the successful operation of web scraping scripts within the Python environment via Airflow, the resultant data was stored in an Amazon S3 bucket. This bucket was configured for public accessibility. To enhance organizational efficiency, distinct folders were created within the bucket, categorized under the labels KAP, Bist Live, and Comments.

3.4 Data Processing

For processing the raw data stored in Amazon S3, Apache Spark and EMR (Elastic MapReduce) were employed. Leveraging Spark within EMR facilitated rapid data manipulation. Structured data, such as the BIST100 live prices, were handled using the DataFrame format. Conversely, unstructured data, including KAP notifications and comments, were managed using the RDD format. The subsequent Spark transformation steps are shown as lineage graph for the comment RDD.

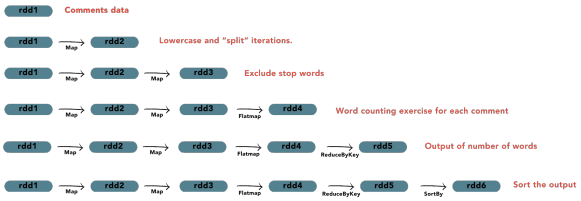


Figure 2. Comments RDD Lineage Graph

Post-processing the data in EMR, we generated Parquet-formatted output for each data source, resulting in a 30% reduction in size.

Once all files were converted to Parquet format and saved back to the S3 bucket, and Apache NiFi was implemented to ensure a seamless flow of data from S3 to Elasticsearch where visualization process can takes place.

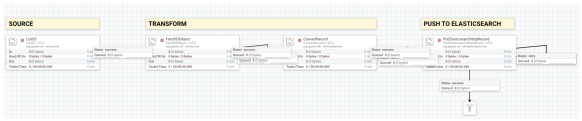


Figure 3. Ni-Fi Flow

Within Apache NiFi, Parquet-formatted data was collected and converted to JSON format to enable storage and retrieval by Elasticsearch. Subsequently, upon pushing the data to Elasticsearch, visualization of the datasets became possible in Kibana, the commonly preferred visualization tool.

3.5 Data Visualization

In terms of data visualization, Kibana was employed to gather and present the datasets processed earlier. The objective was to ensure the dashboard's user-friendliness and goal orientation. The available datasets include:

- BIST100 prices on an hourly/daily basis
- Daily KAP notifications for BIST100
- Daily comments data for BIST100

The aim is to empower customers to utilize RUMOR's dashboard for various purposes. For example, they should be able to track the price movements of specific stocks over time, explore the details of KAP notifications, and understand their impact on prices. Additionally, users should be able to identify trending stocks based on discussions and conduct detailed analyses to make informed decisions. The figures shown below represent analysis-looking graphs generated in Kibana.

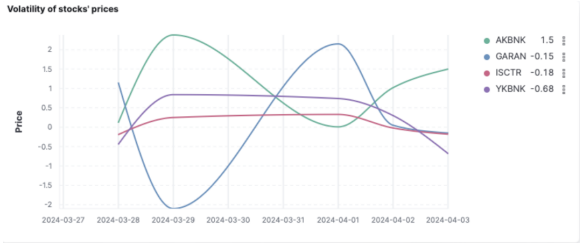


Figure 4. Volatility of stocks' prices

By considering the difference between the previous day's closing price and the current day's opening price, an attempt was made to measure the distance to relative stability. Four stocks in the banking sector were selected from BIST100 as examples showing in Figure 4.

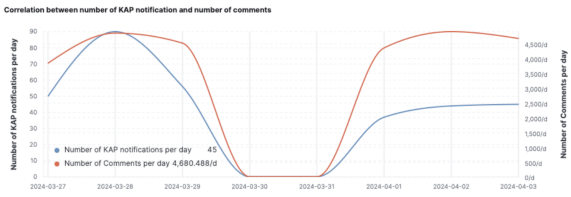
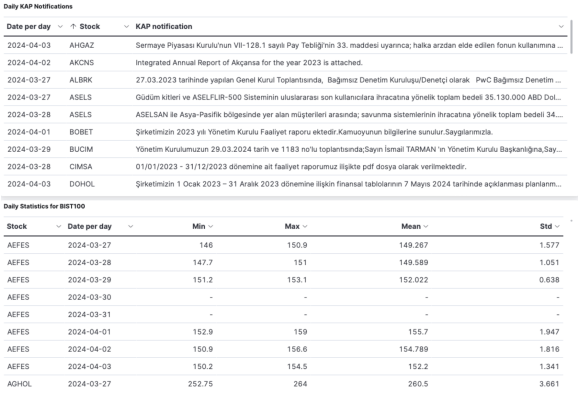


Figure 5. Correlation between number of KAP notifications and comments

The number of KAP notifications and comments were analyzed by standardizing their units to investigate any potential correlations. This involved assessing whether an increase in KAP notifications received on a given day corresponds to a rise in comments made regarding that specific stock. As depicted in Figure 5, there appears to be a positive correlation between them.

Figure 6. Word Cloud

The number of word occurrences has been tracked for each comment, and the output of this data was used to generate a word cloud graph. Figure 6 shows that the most frequent words are the following: 'hisse', 'ytd', 'sonra', 'tavan', 'bugün', etc.



Daily KAP Notifications

Date per day	Stock	KAP notification
2024-04-03	AHGAZ	Sermaye Piyasası Kurulu'nun VII-128.1 sayılı Pay Tebliği'nin 33. maddesi uyarınca; halka arzdan elde edilen fonun kullanımına ...
2024-04-02	AKCNS	Integrated Annual Report of Akçansa for the year 2023 is attached.
2024-03-27	ALBRK	27.03.2023 tarihinde yapılan Genel Kurul Toplantısında, Bağımsız Denetim Kuruluşu/Denetçi olarak PwC Bağımsız Denetim ...
2024-03-27	ASELS	Güdüm kitleri ve ASELFLIR-500 Sisteminin uluslararası son kullanıcılara ihracatına yönelik toplam bedeli 35.130.000 ABD Dol...
2024-03-28	ASELS	ASELSAN ile Asya-Pasifik bölgesinde yer alan müşterileri arasında; savunma sistemlerinin ihracatına yönelik toplam bedeli 34....
2024-04-01	BOBET	Şirketimizin 2023 yılı Yönetim Kurulu Faaliyet raporu ektedir.Kamuoyunun bilgilerine sunulur.Saygılarımızla.
2024-03-29	BUCIM	Yönetim Kurulumuzun 29.03.2024 tarih ve 1183 no'lu toplantısında;Sayın İsmail TARMAN 'ın Yönetim Kurulu Başkanlığına,Say...
2024-03-28	CIMSA	01/01/2023 - 31/12/2023 dönemine ait faaliyet raporumuz ilişikte pdf dosya olarak verilmektedir.
2024-04-03	DOHOL	Şirketimizin 1 Ocak 2023 – 31 Aralık 2023 dönemine ilişkin finansal tablolarının 7 Mayıs 2024 tarihinde açıklanması planlanm...

Daily Statistics for BIST100

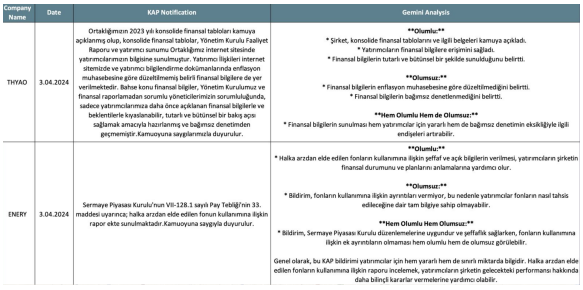
Stock	Date per day	Min	Max	Mean	Std
AEFES	2024-03-27	146	150.9	149.267	1.577
AEFES	2024-03-28	147.7	151	149.589	1.051
AEFES	2024-03-29	151.2	153.1	152.022	0.638
AEFES	2024-03-30	-	-	-	-
AEFES	2024-03-31	-	-	-	-
AEFES	2024-04-01	152.9	159	155.7	1.947
AEFES	2024-04-02	150.9	156.6	154.789	1.816
AEFES	2024-04-03	150.2	154.5	152.2	1.341
AGHOL	2024-03-27	252.75	264	260.5	3.661

Figure 7. Snapshot Information in a Tabular Form

In addition to the visual graphics, snapshot information in tabular form also appears on the dashboard showing in Figure 7, allowing users to scroll, search, or explore the stocks they are interested in. Filters will be available for customer's convenience.

3.6 AI Analysis

One of the most important features of Rumor is its ability to interpret KAP notifications, which can be extremely beneficial, especially given that they can easily seem confusing for people lacking expertise in that area. Rumor's AI support is prepared to explain those complex events in positive, negative, and both positive and negative contexts.



Company Name	Date	KAP Notification	Gemini Analysis
THYAO	3.04.2024	Ortaklığımızın 2023 yılı konsolide finansal tabloları kamuya açıklanmış olup, konsolide finansal tablolar, Yönetim Kurulu Faaliyet Raporu ve yatırımcı sunumu Ortaklığımız internet sitesinde yatırımcılarımızın bilgisine sunulmuştur. Yatırımcı İlişkileri internet sitemizde ve yatırımcı bilgilendirme dokümanlarında enflasyon muhasebesine göre düzeltilmemiş belirli finansal bilgilere de yer verilmektedir. Bahse konu finansal bilgiler, Yönetim Kurulumuz ve finansal raporlamadan sorumlu yöneticilerimizin sorumluluğunda, sadece yatırımcılarımıza daha önce açıklanan finansal bilgilerle ve beklentilerle kıyaslanabilir, tutarlı ve bütünsel bir bakış açısı sağlamak amacıyla hazırlanmış ve bağımsız denetimden geçmemiştir.Kamuoyuna saygılarımızla duyurulur.	**Olumlu:** * Şirket, konsolide finansal tablolarını ve ilgili belgeleri kamuya açıkladı. * Yatırımcıların finansal bilgilere erişimini sağladı. * Finansal bilgilerin tutarlı ve bütünsel bir şekilde sunulduğunu belirtti. **Olumsuz:** * Finansal bilgilerin enflasyon muhasebesine göre düzeltilmediğini belirtti. * Finansal bilgilerin bağımsız denetlenmediğini belirtti. **Hem Olumlu Hem de Olumsuz:** * Finansal bilgilerin sunulması hem yatırımcılar için yararlı hem de bağımsız denetimin eksikliğiyle ilgili endişeleri artırabilir.
ENERY	3.04.2024	Sermaye Piyasası Kurulu'nun VII-128.1 sayılı Pay Tebliği'nin 33. maddesi uyarınca; halka arzdan elde edilen fonun kullanımına ilişkin rapor ekte sunulmaktadır.Kamuoyuna saygıyla duyurulur.	**Olumlu:** * Halka arzdan elde edilen fonların kullanımına ilişkin şeffaf ve açık bilgilerin verilmesi, yatırımcıların şirketin finansal durumunu ve planlarını anlamalarına yardımcı olur. **Olumsuz:** * Bildirim, fonların kullanımına ilişkin ayrıntıları vermiyor, bu nedenle yatırımcılar fonların nasıl tahsis edileceğine dair tam bilgiye sahip olmayabilir. **Hem Olumlu Hem Olumsuz:** * Bildirim, Sermaye Piyasası Kurulu düzenlemelerine uygundur ve şeffaflık sağlarken, fonların kullanımına ilişkin ek ayrıntıların olmaması hem olumlu hem de olumsuz görülebilir. Genel olarak, bu KAP bildirimi yatırımcılar için hem yararlı hem de sınırlı miktarda bilgidir. Halka arzdan elde edilen fonların kullanımına ilişkin raporu incelemek, yatırımcıların şirketin gelecekteki performansı hakkında daha bilinçli kararlar vermelerine yardımcı olabilir.

Figure 8. AI-based Interpretation

The KAP notification showing in Figure 8 from THAYO indicates the public disclosure of their 2023 consolidated financial statements. Gemini Analysis acknowledges the positive aspects of providing access to financial information but raises concerns about the lack of adjustment for inflation accounting and independent audit.

Conclusion

RUMOR is a data-driven platform designed to empower users with insights into the daily trends of publicly traded stocks. It achieves this by combining a multifaceted approach to data collection and analysis:

Comprehensive Data Gathering: RUMOR incorporates real-time and historical stock price movements, news announcements from the Public Disclosure Platform (KAP), and social media commentary to provide a holistic view of market activity.

Automated Workflows: Airflow orchestrates the entire data pipeline, ensuring the seamless flow of information from various sources into a centralized storage system (Amazon S3).

Scalable Storage and Processing: RUMOR leverages the power of Amazon's cloud infrastructure – S3 for secure and scalable data storage, and EMR (Elastic MapReduce) for efficient data processing in a distributed manner. Additionally, the platform utilizes Apache Spark for data transformations, optimizing data for further analysis.

AI-powered Insights: RUMOR employs AI algorithms to analyze KAP notifications, generating insights that highlight both positive and negative aspects of the news, aiding users in making informed decisions.

Intuitive Visualization: The processed data is transferred to Elastic Search for efficient querying and exploration. Kibana, a powerful visualization tool, allows users to create interactive dashboards that display stock price movements, social media sentiment, and news trends, offering a clear picture of market dynamics.

In essence, RUMOR bridges the gap between complex financial data and user comprehension by harnessing the power of data collection, automation, scalable cloud infrastructure, and user-friendly data visualization tools.

References

[1] Ioan Dzitac. Advanced ai techniques for web mining. *IT Department, Agora University*, 2023.

[2] The Economist. Stockmarkets are booming. but the good times are unlikely to last. *Finance and Economics*, February 2022.

[3] Dattatray P. Gandhmal and K. Kumar. Systematic analysis and review of stock market prediction techniques. *Procedia Computer Science*, 216:96–102, March 2023.

[4] KAP. Bist100 stock market data. *KAP*, March 2024.

[5] Ruihua Song and Microsoft Research. Joint optimization of wrapper generation and template detection. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 894, September 14 2007. Archived from the original (PDF) on October 11, 2016.