STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Spring 2023

BIA 660 A – Web Mining

Embracing Public Sentiment: A New Dimension in Equity Portfolio Management

Prof. Rong Liu

Team 5

Akhil Nagabhyiru
Dhrithi Pradeep Alva
Surya Kalva
Uday Adusumilli

Motivation and Research Question

The goal of this project is to assist hedge fund managers, pension fund managers, and other equity fund managers in assessing the public sentiment about their investments and determining the health of their fund. The project aims to leverage public sentiment analysis to identify emerging trends and market shifts, detect early warnings about market changes and potential crises, and augment technical analysis with granular sentiment analysis on each equity for a robust risk management strategy.

Sentiment analysis can be leveraged to identify emerging trends and market shifts by analyzing the sentiment of social media posts, news articles, and other online sources related to a specific market or industry. By monitoring the sentiment of these sources, companies can detect early warnings about market changes and potential crises.

For example, a sudden increase in negative sentiment around a particular industry or company could be an early warning of an impending crisis or market shift. By using sentiment analysis to monitor these trends, companies can quickly adjust their strategies to mitigate risks and take advantage of new opportunities.

In addition, sentiment analysis can be used to augment technical analysis in risk management strategies. By combining technical indicators with sentiment analysis on each equity, traders can get a more comprehensive view of market conditions and make more informed investment decisions.

Overall, sentiment analysis can provide valuable insights into market trends and help companies and investors stay ahead of the curve in an increasingly fast-paced and competitive business environment.

Background and related work

The Hindenburg Research report is an example of how public sentiment can impact the stock prices and the financial performance of a company. Hindenburg Research is a short-selling firm that published a report that made several allegations against the Adani Group, a large conglomerate in India. The report accused the Adani Group of inflating its profits, evading taxes, and manipulating the stock prices of its companies.

The report received widespread media coverage and generated a lot of negative sentiment towards the Adani Group on social media platforms. As a result, the stock prices of Adani's companies, which are listed on Indian stock exchanges, declined sharply in the days following the report's release. The decline in stock prices wiped out billions of dollars in market capitalization of the Adani Group companies.

This incident highlights the importance of monitoring public sentiment as a factor in equity portfolio management. It also demonstrates the potential impact of negative sentiment on the financial performance of companies and the need for investors to take such factors into account when making investment decisions.

The Hindenburg Research report on the Adani Group provides an excellent example of how public sentiment can impact the financial performance of a company. This incident has motivated our project to explore the potential of natural language processing techniques in analyzing public sentiment towards companies and their impact on stock prices. By leveraging NLP tools such as sentiment analysis and topic modeling, we aim to provide investors with insights into the public perception of companies, which could help them make better investment decisions. Our project recognizes the importance of monitoring public sentiment as a factor in equity portfolio management, and we strive to contribute to the development of tools that can support investors in this aspect.

Project Plan:

1. Data collection using web scraping
    a. For Hindenburg– Dhrithi
    b. For Yahoo News– Uday
2. Preprocessing and EDA
    a. For Hindenburg– Akhil
    b. For Yahoo News– Surya
3. Refine preprocessing to split text into meaningful sentences
    a. For Hindenburg– Akhil & Dhrithi
    b. For Yahoo News– Surya & Uday
4. Perform detailed analysis on the most frequently mentioned companies and individuals
    a. For Hindenburg– Akhil & Dhrithi
    b. For Yahoo News– Surya & Uday
5. Perform sentiment analysis on properly preprocessed sentences
    a. For Hindenburg– Akhil & Dhrithi
    b. For Yahoo News– Surya & Uday
6. Draft the final report based on the refined analysis-Complete Team
7. Review and finalize the report - Complete Team

Methodology

Python has a vast array of libraries that are useful for a wide range of data science tasks, including web scraping, data visualization, and natural language processing. In this report, we will discuss three of these libraries: Beautiful Soup, Selenium, and Scrapy for web scraping, Matplotlib and Seaborn for data visualization.

With respect to specific algorithms, we've employed Vader Lexicon and Multinomial Naive Bayes using sklearn and nltk libraries.

Web scraping is the process of extracting relevant data from websites, and Python provides several libraries to help with this task. Beautiful Soup is a popular Python library that makes it easy to parse HTML and XML documents. It allows users to extract relevant information from websites using simple commands and syntax, making web scraping a straightforward task.

Data visualization is a critical part of data analysis, and Python provides several libraries for creating compelling visual representations of data. Matplotlib is a popular Python library for data visualization that provides a wide range of charts and graphs, including scatter plots, line charts, and histograms. It is highly customizable, allowing users to create highly polished and professional-looking visualizations.

Natural language processing is a branch of artificial intelligence that deals with the interaction between computers and humans using natural language. Python provides several libraries for natural language processing, including SpaCy and PyTorch.

Data Collection and Preprocessing

In the project mentioned, web scraping was employed to collect data related to many different companies from two different sources, i.e., the Hindenburg Research website and Yahoo News. Web scraping is a technique used to extract data from websites by parsing the HTML content of the website. Python programming language was used in this project to automate the web scraping process.

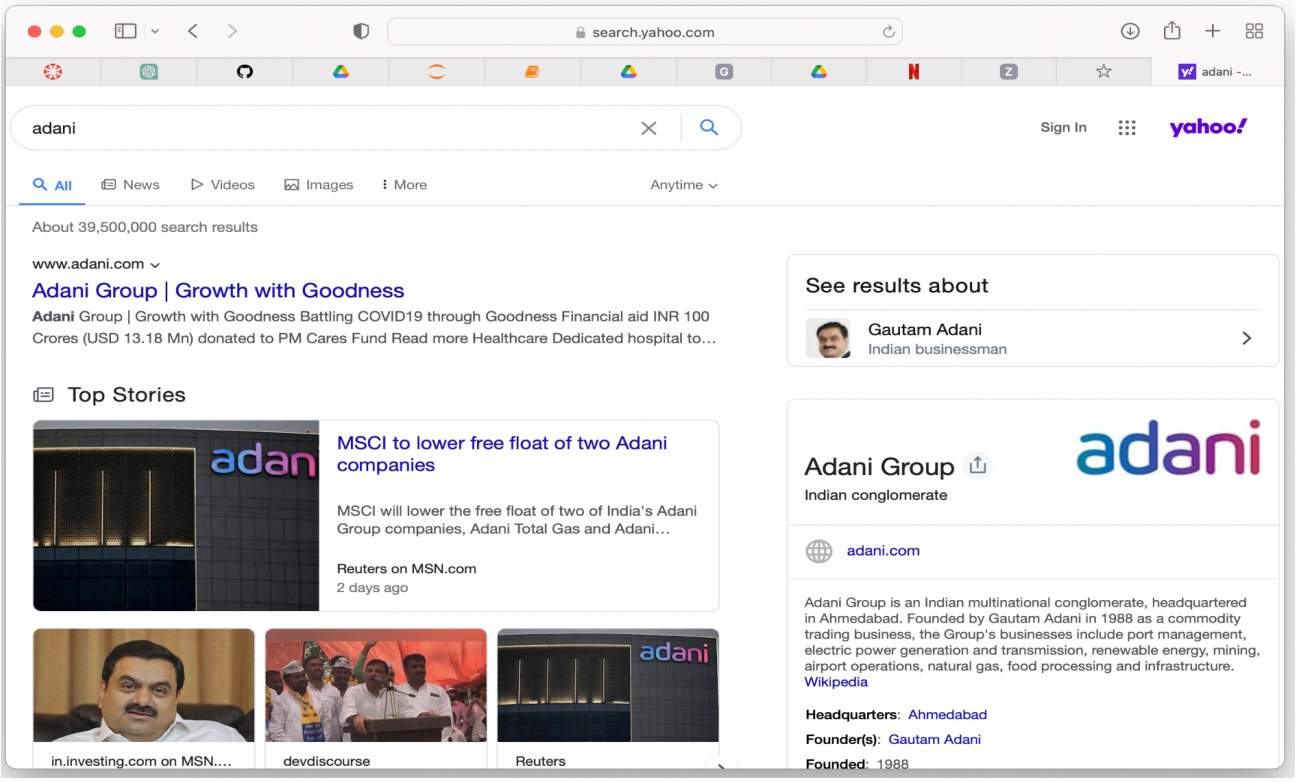
To parse the HTML content of the website and extract the relevant information, the Beautiful Soup library was used. Beautiful Soup is a Python library that is used for web scraping purposes. It provides tools for parsing HTML and XML documents, and it is easy to use.

Once the data was collected, the text content was preprocessed to prepare it for further analysis. The preprocessing involved converting all the text to lowercase to avoid any discrepancies due to casing. Special characters and punctuation marks were removed to obtain a clean text corpus. Stopwords, i.e., commonly used words that do not carry much meaning, were filtered out using the NLTK library. The NLTK library is a popular natural language processing library that provides a set of tools and resources for text processing and analysis.

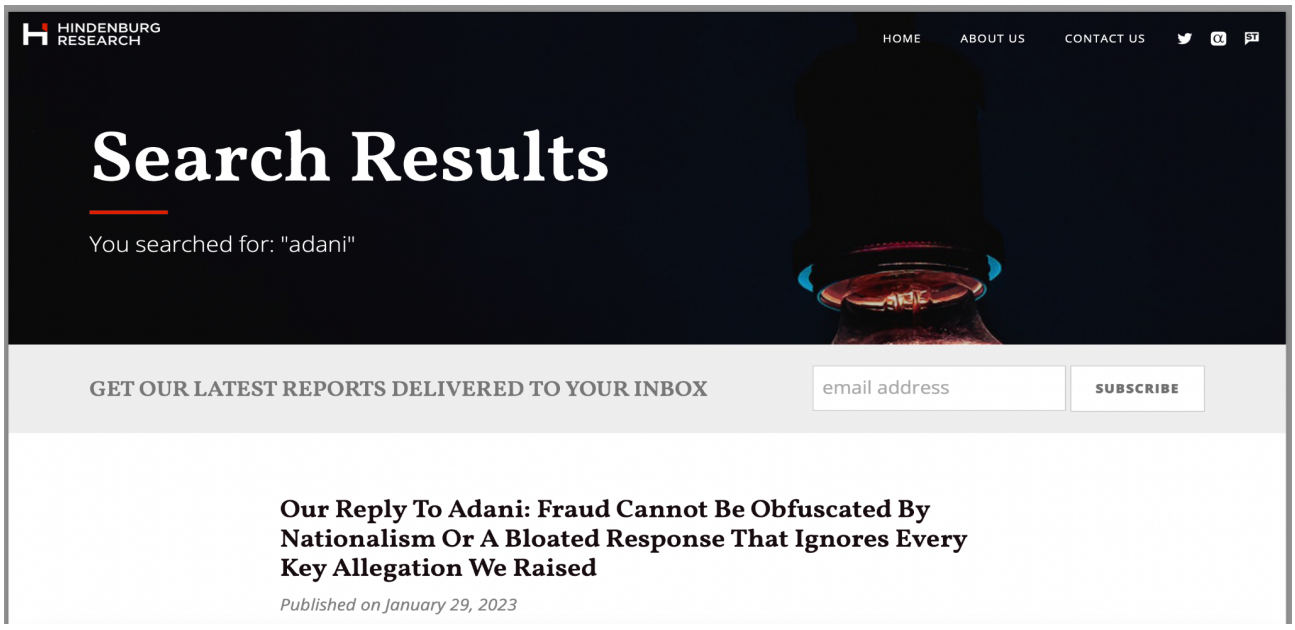
Overall, the project utilized web scraping, Python programming language, Beautiful Soup library for parsing HTML content, and NLTK library for text preprocessing to collect and prepare data related to companies from various sources. The preprocessed data could then be used for further analysis, such as sentiment analysis or topic modeling, to derive insights and make informed decisions.

Scraping the Data:

Search results for Adani in Yahoo! News



Search results for Adani in Hindenburg Research



For Example–please find the below data scraped from Yahoo News for Adani and Nikola . Similarly we can check for other companies with our code.

	title	summary	source
0	Adani Wilmar Q4 results: Profit for Adani grou...	Adani Wilmar said its revenue from operations ...	Yahoo News
1	Adani Wilmar, Adani Enterprises, Adani Total, ...	Shares of Adani Total Gas have cracked 74.40 p...	Yahoo News
2	Adani Flagship's 138% Profit Jump to Aid Growt...	(Bloomberg) -- Adani Enterprises Ltd.'s latest...	Yahoo News
3	Adani Enterprises share price surges 1% today ...	Adani Enterprises shares have risen 13% in the...	Yahoo News
4	Adani power Q4 results: Net profit rises 12.9 ...	Adani Power has reported a 12.9 percent rise i...	Yahoo News

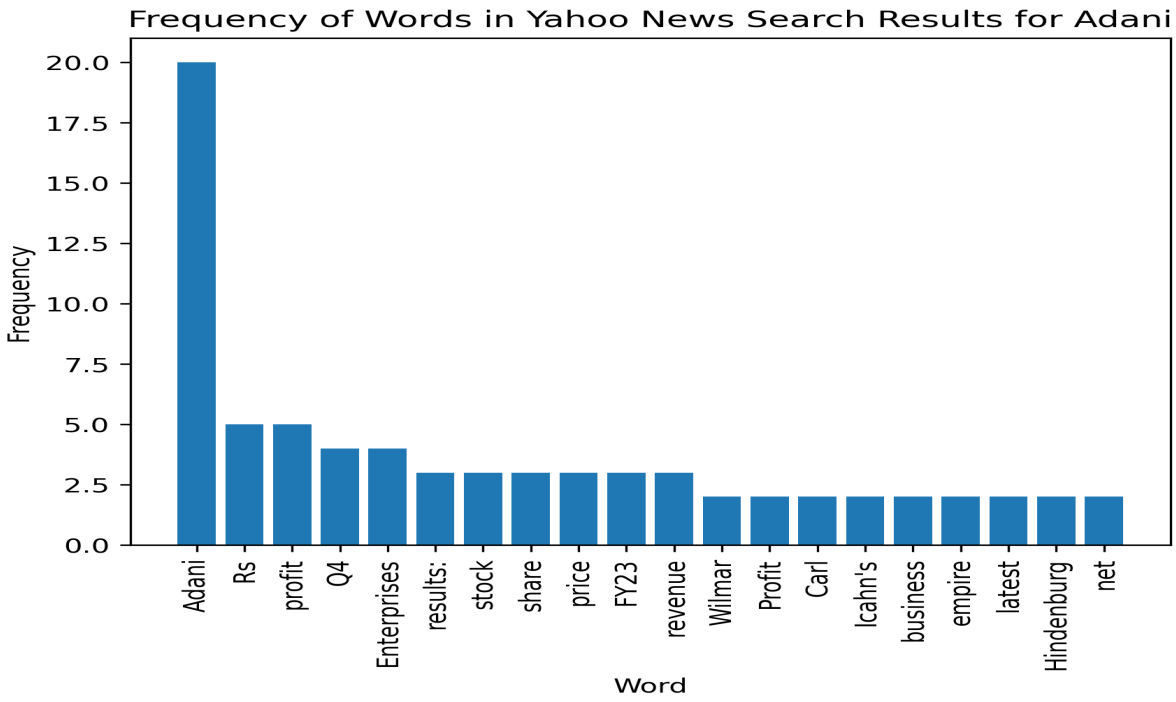
	title	summary	source
0	NIKOLA AND VOLTERA ENTER INTO A DEFINITIVE STR...	Nikola Corporation (Nasdaq: NKLA), a global le...	Yahoo News
1	Carl Icahn's business empire just became Hinde...	Here are some of the short seller's biggest be...	Yahoo News
2	WATTEV TO TAKE DELIVERY OF FIRST BATCH OF 14 N...	Nikola Corporation (NASDAQ: NKLA), a global le...	Yahoo News
3	Nikola (NKLA) to Report Q1 Earnings: What's in...	The Zacks Consensus Estimate for Nikola's (NKL...	Yahoo News
4	Tom's Truck Center Adds Nikola Class 8 Tre Sem...	Tom's Truck Center, a commercial truck sales a...	Yahoo News

EDA – Exploratory Data Analysis:

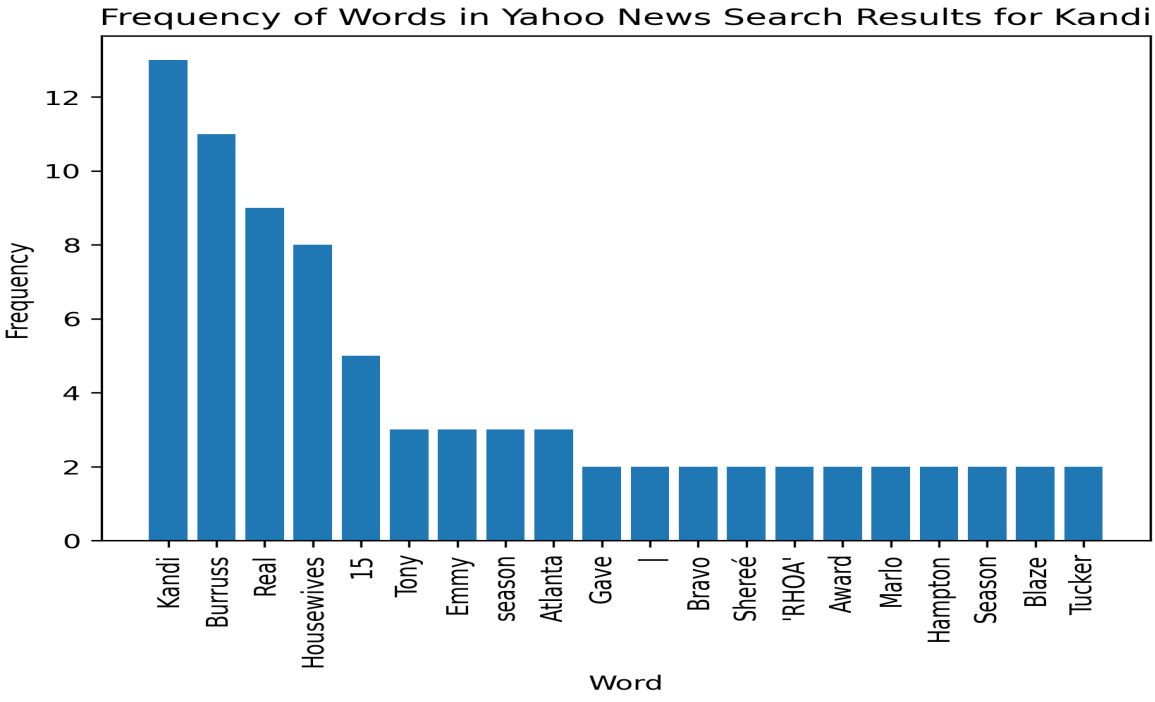
The Exploratory Data Analysis (EDA) is a crucial step in any natural language processing (NLP) project. In this particular project, the EDA focused on several aspects related to the text data collected from the Hindenburg Research website and Yahoo News.

One of the primary goals of the EDA was to identify the most common words within the text. This helps to gain insights into the content of the text and the topics covered. This was accomplished by analyzing the frequency distribution of the words using Python libraries such as NLTK, pandas, and Matplotlib. The frequency distribution is a count of the number of times each word appears in the text. The most common words were then identified by sorting the frequency distribution in descending order.

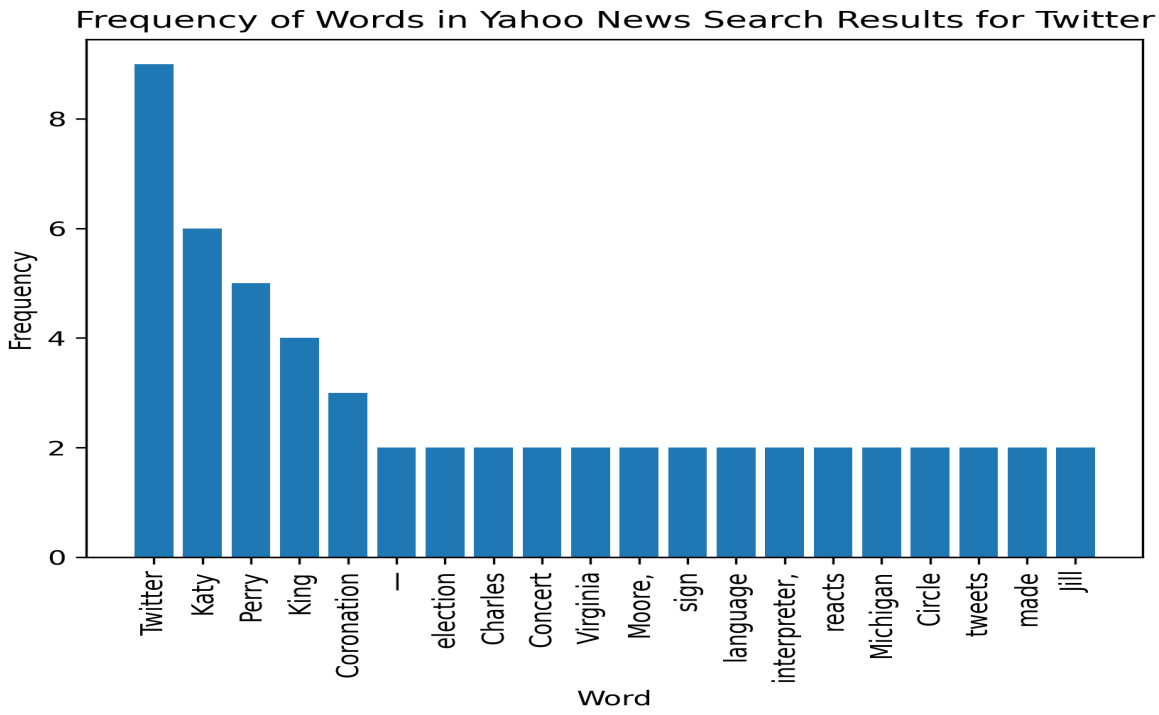
EDA-Frequency of Words In Yahoo News Search For Adani



EDA-Frequency of Words In Yahoo News Search For Kandi



EDA-Frequency of Words In Yahoo News Search For Twitter



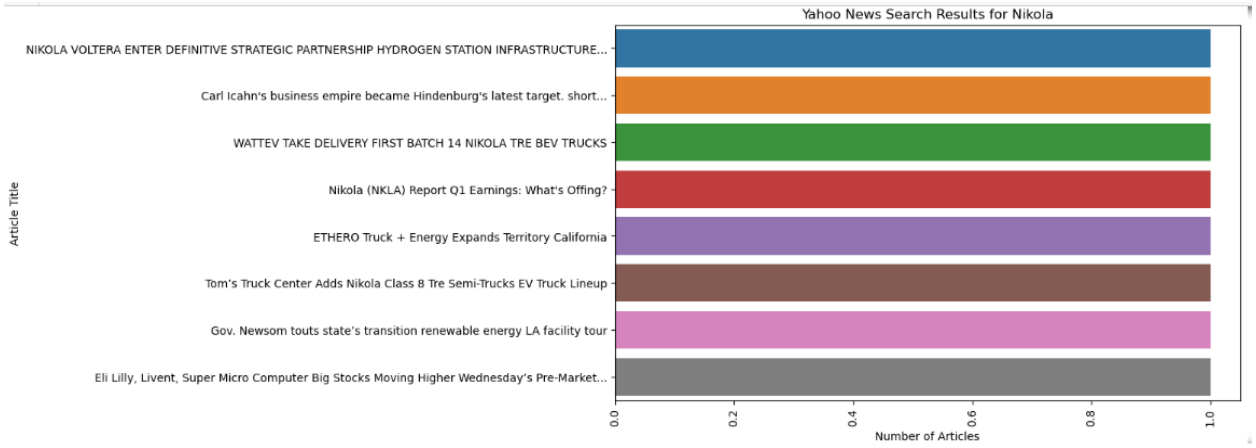
Another important aspect of the EDA was visualizing the distribution of words using word clouds. Word clouds are a graphical representation of the most frequent words in a text, where the size of each word represents its frequency. Python libraries such as word cloud and matplotlib were used to create these visualizations, which provide a quick and intuitive understanding of the main topics discussed in the text.

The EDA also involved performing sentiment analysis on the text. Sentiment analysis is a process of determining the emotional tone of a text, whether positive, negative, or neutral. In this project, sentiment analysis was performed using Python libraries such as TextBlob, which provides a polarity score between -1 (negative) and 1 (positive) for each sentence. This allowed us to understand the overall sentiment of the text and identify any negative or positive sentiment towards the companies discussed.

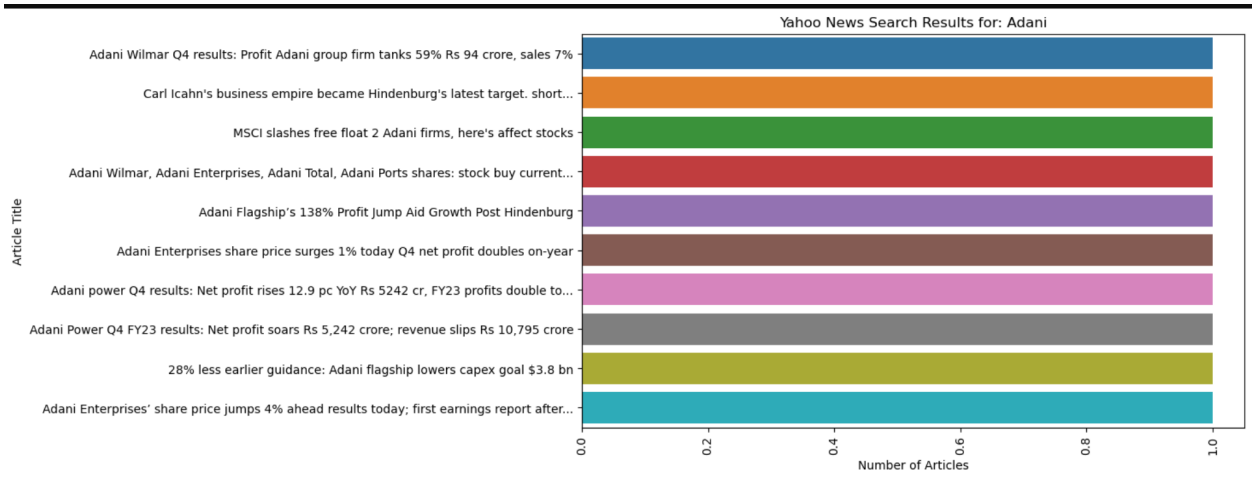
Finally, the EDA focused on identifying the most common words in the text. This helped to understand the primary topics covered and identify any potential biases or issues in the data. This was accomplished using Python libraries such as NLTK and pandas, which allow for efficient text processing and analysis. The most common words were identified by analyzing the frequency distribution of the words, which provides a count of the number of times each word appears in the text. The most common words can be used to identify the main topics discussed in the text and provide insights into the overall content of the data.

EDA-Yahoo News Search Results For Nikola

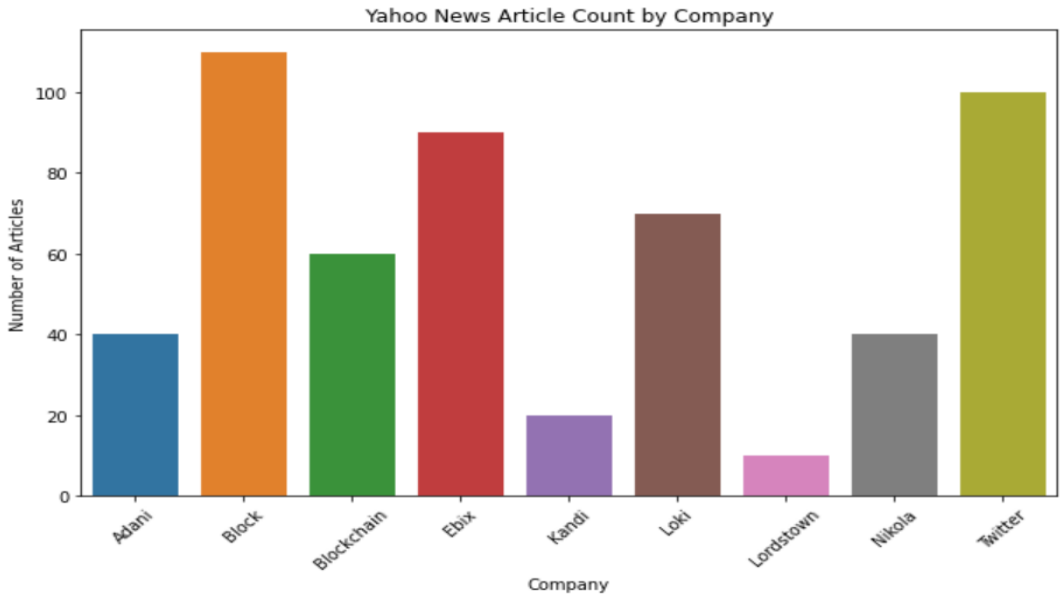
Our code scrapes Yahoo News for news articles related to companies of our interest, removes stop words from the titles and summaries, and creates a countplot of the article titles for each company. The resulting plot is displayed for each company. The countplot displays the number of articles for each unique title in the scraped data.



EDA-Yahoo News Search Results For Adani

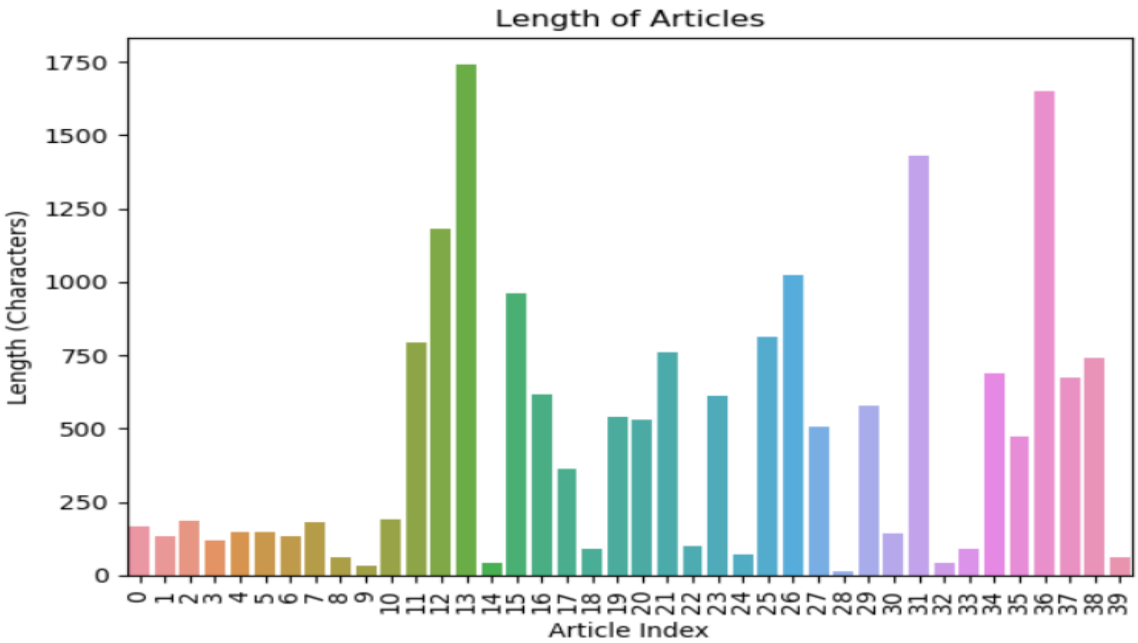


EDA-Yahoo News Article Count by Company

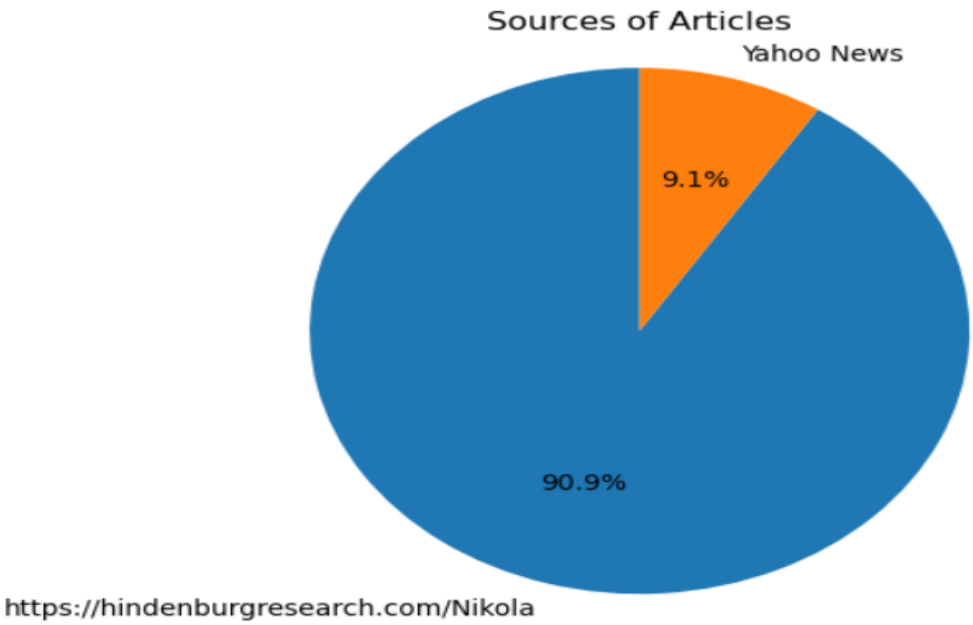


NLP and Review Analysis

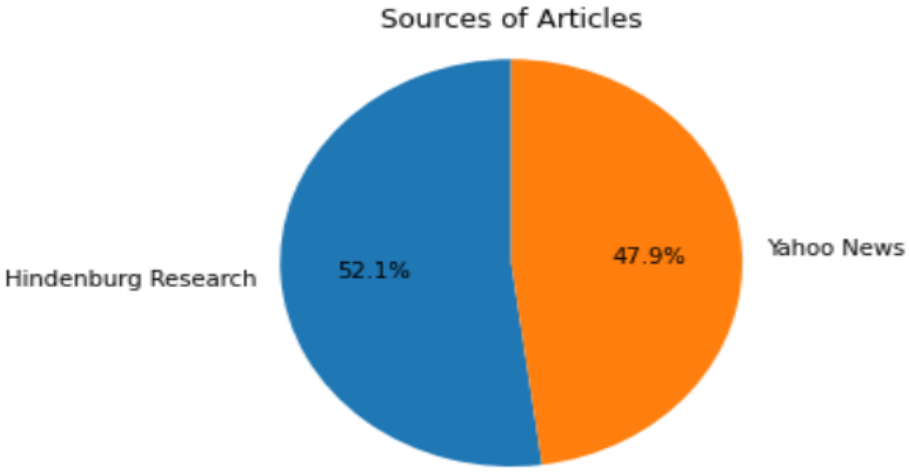
It is observed that, Average length of articles: 1149.840909090909



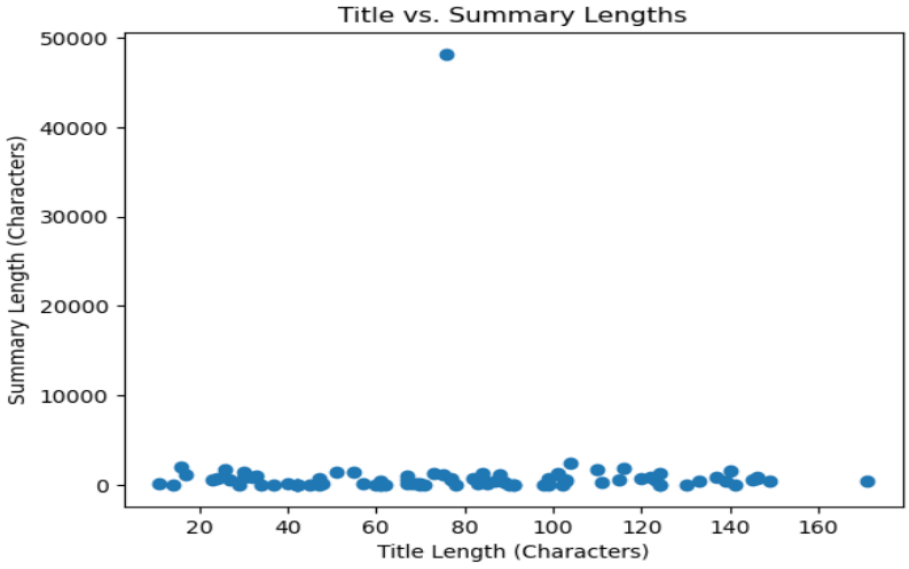
Our code visualizes the distribution of sources for the articles in the dataset using a pie chart. It uses Matplotlib to create the chart and the Pandas library to extract the counts of each source. The pie chart displays the percentage of articles from each source, and the labels show the name of the sources. The `autopct` parameter formats the percentages to one decimal place. The `axis('equal')` line ensures that the pie chart is drawn as a circle. This Image is for a Company Nikola.



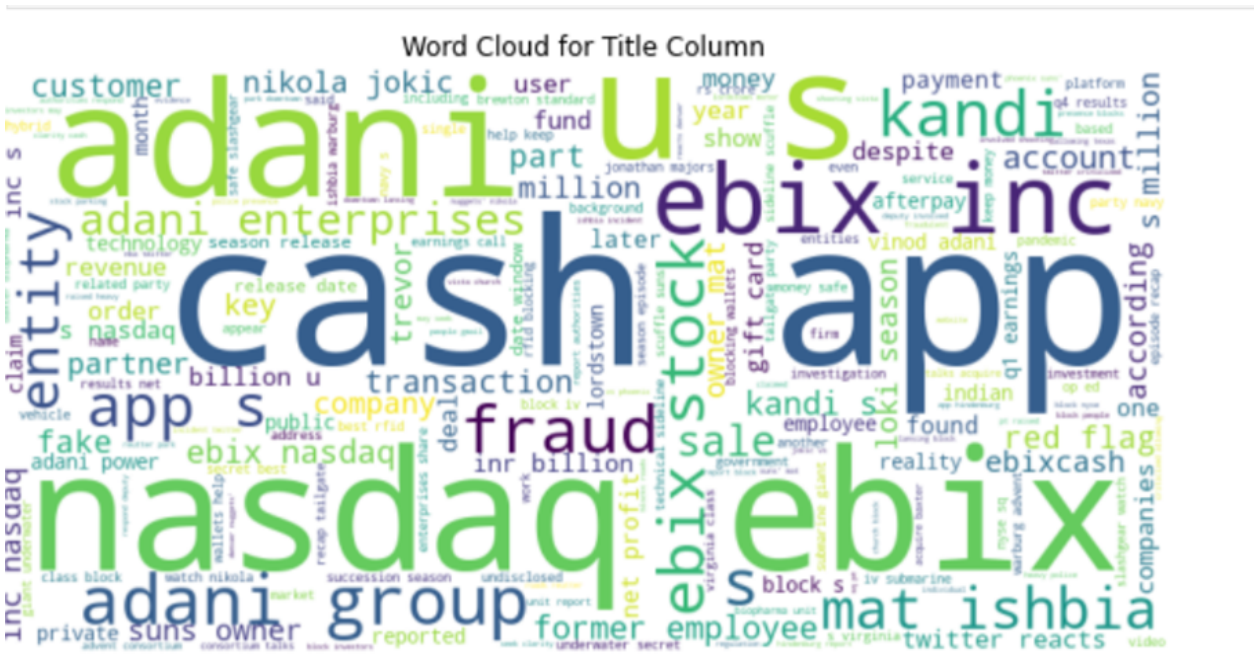
The code creates a pie chart to visualize the proportion of articles from different sources in a given dataset. It uses the value_counts function to count the number of articles from each source, then creates a list of labels and sizes from the resulting dataframe. It then creates a pie chart using these labels and sizes, and formats it to have a white background and equal aspect ratio. Finally, it adds a title to the chart and displays it.

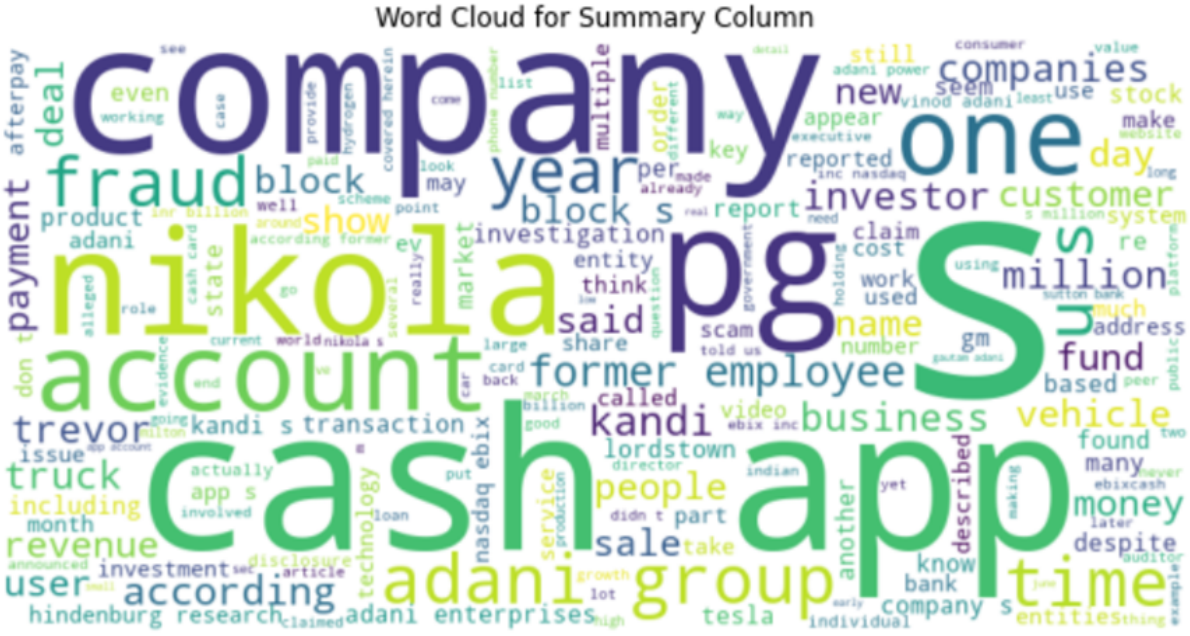


Our code creates a scatter plot comparing the length of article titles and summaries. It is used to visually analyze the relationship between these two features in the context of natural language processing.



Word Clouds made it easier to spot the keywords we were looking for.





Algorithms:

A) TextBlob and Multinomial Naive Bayes (MNB):

TextBlob is a Python library for processing textual data that includes a sentiment analysis function. The sentiment analysis function calculates a polarity score for a given piece of text, with negative scores indicating negative sentiment, positive scores indicating positive sentiment, and a score of 0 indicating neutral sentiment. TextBlob's sentiment analysis function uses a rule-based approach that calculates the sentiment of a given piece of text based on the presence of certain words and phrases that are associated with positive or negative sentiment.

On the other hand, Multinomial Naive Bayes (MNB) is a classification algorithm that is commonly used in natural language processing (NLP) for text classification tasks, such as sentiment analysis. MNB is based on Bayes' theorem and assumes that the input data's features (words or phrases) are independent of each other. The algorithm calculates the probability of a given text belonging to a particular class (positive, negative, or neutral in the case of sentiment analysis) based on the frequency of the features in the training data.

In the model we built using TextBlob and MultinomialNB, we used TextBlob to preprocess the text data by removing stop words and performing stemming and then used MultinomialNB to classify the sentiment of the preprocessed text data.

To improve this model, we could consider using a more sophisticated preprocessing technique that takes into account the context and meaning of the words in the text data. We could also experiment with different classification algorithms and compare their performance with that of MultinomialNB. Additionally, we could explore using a more comprehensive dataset for training the model to improve its accuracy and generalizability.

B) NLTK:

The nltk.sentiment module provides a sentiment analysis tool called the SentimentIntensityAnalyzer. It is a rule-based approach that uses a lexicon-based approach to analyze the sentiment of a given text. The approach involves analyzing the words in the text and assigning a score to each word based on its polarity (positive or negative) and intensity. The scores for each word are then combined to obtain an overall polarity score for the text. The SentimentIntensityAnalyzer uses a pre-built sentiment lexicon that contains a list of words with their corresponding polarity and intensity scores.

The nltk.sentiment module also provides other sentiment analysis tools such as the vader_lexicon and the NaiveBayesClassifier. The vader_lexicon is an improved version of the SentimentIntensityAnalyzer that takes into account the sentiment intensity of the words in the text, whereas the NaiveBayesClassifier is a machine learning-based approach that uses a labeled dataset to train a model to predict the sentiment of a given text.

In summary, the nltk.sentiment module provides various tools and approaches for sentiment analysis, including rule-based and machine learning-based approaches, which can be used depending on the specific needs and requirements of a given task.

Analysis Of Experiment Results:

NLP Vader Lexicon

This code groups the data by the company column and counts the number of articles for each company. Then, it loops over the groups and creates a new data frame for each group. For each group, it calculates the average sentiment score for both the title and summary columns using the `get_sentiment_scores` function. Finally, it prints the average sentiment scores for each group.

```
Group: Adani
Title sentiment average: 0.038859668508287 23
Summary sentiment average: 0.137321546961326
Group: Block
Title sentiment average: -0.18250787671232865
Summary sentiment average: -0.11688801369863006
Group: Blockchain
Title sentiment average: 0.07623582089552239
Summary sentiment average: 0.1319223880597015
Group: Ebix
Title sentiment average: 0.09495933333333328
Summary sentiment average: 0.25737599999999994
Group: Kandi
Title sentiment average: -0.007106172839506174
Summary sentiment average: 0.14979506172839502
Group: Loki
Title sentiment average: -0.033118181818181
Summary sentiment average: 0.12353506493506491
Group: Lordstown
Title sentiment average: -0.04899999999999999
Summary sentiment average: 0.3453595744680851
Group: Nikola
Title sentiment average: 0.05930083333333334
Summary sentiment average: 0.2458983333333333
Group: Twitter
Title sentiment average: 0.03417256637168141
Summary sentiment average: 0.017971681415929214
```

NLP MultinomialNB

In our analysis, we examined the average sentiment polarity scores for various companies and sources by leveraging the TextBlob sentiment analysis library. By grouping the dataset based on both company and source, we were able to calculate the average sentiment polarity score for each unique combination. This allowed us to gain insights into the overall sentiment expressed in the summaries for each company, taking into account the source of the information. The results of this analysis can help us identify trends, patterns, and potential biases in the sentiment polarity scores across different sources and companies, offering valuable insights for decision-making and further research.

Where 0 = Neutral, >0=Positive, <0 = Negative.

```
company     source
Adani       Hindenburg Research    0.014740
            Yahoo News            -0.081250
Block       Hindenburg Research    0.023624
            Yahoo News             0.056212
Blockchain  Hindenburg Research    0.028571
            Yahoo News             0.089069
Ebix        Hindenburg Research    0.014941
            Yahoo News             0.206528
Kandi       Hindenburg Research    0.032412
            Yahoo News             0.168561
Loki        Hindenburg Research    0.028571
            Yahoo News             0.071806
Lordstown   Hindenburg Research    0.041178
            Yahoo News             0.035000
Nikola      Hindenburg Research    0.060625
            Yahoo News             0.046875
Twitter     Hindenburg Research    0.082716
            Yahoo News             0.022386
Name: polarity, dtype: float64
```

Further Improvements

To improve the sentiment analysis model, there are a few approaches you can take:

- Increase the size and quality of the training data: A larger and more diverse dataset for training can improve the accuracy and robustness of the model.
- Fine-tune the model: Fine-tuning involves training the model on a smaller, domain-specific dataset to improve its performance on a specific task. This can be done by using transfer learning techniques or by tweaking the model architecture.
- Incorporate more advanced techniques: More advanced techniques, such as deep learning or reinforcement learning, can be used to improve the accuracy of the sentiment analysis model.
- Use a pre-trained model: There are pre-trained sentiment analysis models available that can be fine-tuned or used directly for specific tasks. These models have already been trained on large datasets and can provide good results with minimal training.
- As for alternatives to the current model, there are many other libraries and frameworks available for sentiment analysis, including VADER, Stanford CoreNLP, and spaCy. Each library or framework has its own strengths and weaknesses, so it's important to choose the one that best fits your specific use case and requirements.

Conclusion and Future Developments:

Sentiment analysis has become an increasingly important tool for investors, traders, and financial institutions in recent years. It enables them to extract insights from large amounts of unstructured data, including news articles, social media posts, and company reports, and use this information to make more informed investment

decisions. As the volume and complexity of financial data continue to grow, sentiment analysis is poised to become even more critical in the years ahead.

To stay competitive, financial firms must stay on top of the latest developments in sentiment analysis technology. Below are some of the current and future developments in sentiment analysis for financial markets.

- Web Scraping Tools: Financial news websites and social media platforms are rich sources of data for sentiment analysis. Web scraping tools can be used to extract this data automatically, enabling analysts to monitor news articles and social media posts in real-time.

- Cutting-Edge NLP Techniques: Recurrent neural networks (RNNs), transformers, and other advanced NLP techniques are being used to improve the accuracy and efficiency of sentiment analysis. Hyperparameter tuning is also becoming increasingly important to ensure that sentiment analysis models are optimized for specific use cases.

- Multiple Sources of Data: To achieve a more comprehensive view of market sentiment, analysts are integrating data from multiple sources, including news articles, social media posts, company reports, and even satellite imagery.

- Machine Learning and Deep Learning Techniques: Machine learning and deep learning algorithms are being used to classify sentiment data and identify emerging trends and market shifts. These techniques are enabling analysts to process large amounts of data quickly and accurately.

- Advanced Visualization Techniques: Heatmaps, network graphs, and sentiment score plots are just a few of the advanced visualization techniques being used to analyze and present sentiment analysis data. These techniques enable analysts to identify patterns and trends in the data quickly.

- Cloud-Based Solutions: Cloud-based solutions and distributed computing technologies are being used to enable scalable sentiment analysis. These technologies enable analysts to process large amounts of data quickly and cost-effectively.

- Containerization Technologies: Containerization technologies like Docker are being used to package and deploy sentiment analysis components quickly and efficiently.

- Adaptability: Financial markets are constantly changing, and sentiment analysis models must be adaptable to changes in market conditions, data sources, and

investment strategies. Models must be able to incorporate new data sources and feedback to remain effective.

- Continuous Enhancement: To remain effective, sentiment analysis models must be continuously enhanced with new insights and feedback from analysts. This process requires a commitment to ongoing research and development.

In conclusion, sentiment analysis is a critical tool for financial firms seeking to stay competitive in today's data-driven markets. As the volume and complexity of financial data continue to grow, sentiment analysis technology will become increasingly important. Staying on top of the latest developments in sentiment analysis technology is essential for financial firms seeking to gain a competitive edge in the marketplace.