

Vivekanand Education Society's Institute of Technology (An
Autonomous Institute Affiliated to University of Mumbai,) (Approved by
A.I.C.T.E and Recognized by Govt. of Maharashtra)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA
SCIENCE



A REPORT
ON
“Stock Sentiment Analysis from news headlines”

B.E. (AIDS)

SUBMITTED BY

Swayam Gaikwad (Roll No. 17)
Khyati Hegde (Roll No. 24)
Anjali Parwani (Roll No. 45)

UNDER THE GUIDANCE OF

PROF. Dr. Anjali Yeole
(Academic Year: 2024-2025)

INDEX

Content	Page No.
1. Introduction	3
1.1 Background and Context	3
1.2 Problem Statement	4
1.3 Objectives	5
2. Literature Review	6
2.1 Overview of Related Works	6
2.2 Summary of Key Findings	7
3. Dataset Description	8
3.1 Dataset Overview	8
3.2 Web Scraping of Data	9
4. Methodology	10
4.1 Data Extraction	10
4.2 Headline Parsing	10
4.3 Sentiment Analysis using NLTK Vader	10
4.4 Data Transformation	10
5. Results and Visualizations	11
5.1 Result	11
5.2 Visualization	12
6. Conclusion	13
6.1 Summary of Key Findings	13
6.2 Conclusion	14
6.3 Future Work	14

1. Introduction

1.1 Background and Context

In the modern era of finance, the rapid dissemination of news through digital platforms has transformed how investors and traders make decisions. Traditionally, financial news spread through radio, newspapers, and word of mouth, often taking days to reach the public. Today, with the advent of the internet, news articles can be generated almost instantaneously from earnings reports and corporate announcements. This shift has enabled hedge funds and individual traders to leverage data science techniques, such as sentiment analysis, to gain insights and potentially profit from the wealth of information available.

This project aims to harness the power of sentiment analysis on financial news headlines sourced from FINVIZ.com, a trusted financial news aggregator. By focusing on headlines, which are typically more concise and consistent in structure compared to full-length articles, we can better parse and analyze the underlying sentiment. This approach allows us to gauge the emotional tone of headlines, offering insights into market perceptions regarding specific stocks.

The project employs natural language processing techniques to parse HTML files containing historical news headlines for prominent companies like Facebook and Tesla. Using the NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis model, we augment the model's lexicon with finance-specific jargon to improve accuracy. Through this analysis, we can predict market sentiment based on news headlines, allowing for educated guesses on stock performance.

By visualizing sentiment trends over time, we aim to provide valuable insights that could inform trading strategies while acknowledging the inherent risks of stock market investments. This notebook serves as a foundational exploration of how sentiment analysis can be integrated into financial decision-making processes, highlighting the increasing importance of data-driven approaches in the investment landscape.

1.2 Problem Statement

Stock sentiment analysis is a critical application of Natural Language Processing (NLP) in the financial industry, where the goal is to determine the sentiment or emotions expressed in news headlines regarding stock markets. Financial news has a significant influence on investor behavior and, consequently, stock prices. Positive news often leads to price increases, while negative headlines may result in a drop in stock value.

In this project, "Stock Sentiment Analysis from News Headlines," we aim to develop a model that can automatically analyze news headlines related to stocks and classify them into different sentiment categories, such as positive, negative, or neutral. By leveraging NLP techniques, this analysis helps investors make informed decisions by providing insights into the overall market mood.

This task involves preprocessing textual data from news headlines, applying sentiment analysis algorithms, and building machine learning models to predict the sentiment. The outcome of this analysis has direct implications for stock market predictions, enabling investors and analysts to gauge market trends more efficiently and proactively respond to potential risks or opportunities.

1.3. Objectives

1. To Develop a Robust Sentiment Analysis Model:

Create a model that accurately classifies the sentiment of news headlines related to stocks into positive, negative, or neutral categories. This involves leveraging various NLP techniques, including traditional machine learning algorithms and advanced deep learning models.

2. To Extract and Preprocess Relevant Data:

Implement a systematic approach to collect and preprocess stock news headlines from various sources. This includes cleaning the text data, removing noise, and ensuring consistency in formatting (e.g., date and time).

3. To Enhance Sentiment Analysis with Domain-Specific Lexicons:

Incorporate and customize financial-specific sentiment lexicons (e.g., Loughran-McDonald) to improve the accuracy of sentiment classification. This will help capture the nuances and specific language used in financial news.

4. To Evaluate the Impact of Sentiment on Stock Prices:

Analyze the correlation between the sentiment of news headlines and subsequent stock price movements. This objective aims to provide insights into how public sentiment as reflected in the news can influence market trends.

5. To Assess the Effectiveness of Various Machine Learning Techniques:

Compare the performance of different machine learning models (e.g., Naive Bayes, SVM, LSTM, BERT) in terms of accuracy, precision, recall, and F1-score. This evaluation will help identify the most effective methods for stock sentiment analysis.

6. To Investigate the Role of Data Preprocessing Techniques:

Examine the effects of various data preprocessing techniques (e.g., stop word removal, stemming, lemmatization) on the performance of sentiment analysis models. This objective aims to identify best practices for preparing text data in the financial context.

7. To Explore Feature Engineering Techniques:

Utilize feature engineering methods, such as combining n-grams with sentiment scores, to enhance the performance of sentiment analysis models. This includes identifying relevant features that significantly contribute to accurate sentiment classification.

2. Literature Review

2.1 Overview of Related Works in Sentiment Analysis

Sentiment analysis, also known as opinion mining, has gained significant attention in recent years due to the proliferation of social media and user-generated content. Researchers have developed various methods and techniques to analyze sentiments expressed in text data. The following are some notable works in this field:

- *Traditional Machine Learning Approaches:*

Early sentiment analysis studies primarily focused on traditional machine learning techniques, such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. For instance, Pang et al. (2002) demonstrated the effectiveness of SVM for sentiment classification tasks, achieving promising results on movie reviews.

- *Lexicon-Based Approaches:*

Some studies have utilized lexicon-based methods, where sentiment scores are assigned to words based on predefined sentiment lexicons (e.g., SentiWordNet). Hu and Liu (2004) proposed a method for opinion mining in product reviews that combined lexicon-based sentiment scoring with aspect extraction to identify sentiments related to specific product features.

- *Deep Learning Techniques:*

Recent advancements in deep learning have led to the use of neural networks for sentiment analysis. Kim (2014) introduced a convolutional neural network (CNN) for sentence classification, which outperformed traditional models on various sentiment analysis benchmarks. Other architectures, such as recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs), have also been employed for sentiment classification tasks.

- *Contextualized Word Embeddings:*

The introduction of models like Word2Vec, GloVe, and BERT has revolutionized sentiment analysis. [Devlin et al. \(2018\)](#) presented BERT (Bidirectional Encoder Representations from Transformers), which captures context in text and has set new benchmarks for various NLP tasks, including sentiment analysis.

2.2 Summary of Key Findings from Other Studies

- *Impact of Preprocessing Techniques:*

Studies show that effective data preprocessing (e.g., removing stop words, stemming, and lemmatization) significantly improves the performance of sentiment classification models.

- *Feature Engineering:*

Research indicates that feature engineering plays a crucial role in sentiment analysis. Studies have demonstrated that using a combination of n-grams and sentiment lexicons can enhance the performance of machine learning classifiers.

- *Ensemble Methods:*

Studies have highlighted the effectiveness of ensemble methods in sentiment analysis, where combining predictions from multiple models leads to improved accuracy and robustness.

- *Application in Various Domains:*

Sentiment analysis has been applied across various domains, including finance, healthcare, and social media. For instance, [Schumaker and Chen \(2009\)](#) explored the use of sentiment analysis in predicting stock market trends based on Twitter data, illustrating its practical implications.

3. Dataset

3.1 Dataset Overview

The dataset used in this project focuses on financial news articles, specifically targeting Tesla (TSLA) and other stock-related news. This data was collected through web scraping from Finviz and other financial news sources. The dataset provides insights into market trends, company performance, and how various news events influence stock prices.

1. News Articles

Format: HTML Files

Content: The dataset comprises financial news articles collected from multiple sources such as financial news websites, stock market analysis platforms, and company blogs. These articles provide valuable insights into market trends, company performance, stock movements, and economic developments.

Topics: The news articles are categorized into the following topics:

- Stock performance updates
- Earnings reports and financial results
- Market predictions and analyst opinions
- Macroeconomic news impacting the overall market

2. Categories

The articles are grouped based on the nature of the news:

Stock Performance: Articles covering stock price changes, trading volume, and investor actions.

Earnings: News related to company earnings reports, revenue updates, and profit projections.

Market Predictions: Expert analyses and forecasts regarding market movements or specific stocks.

Macroeconomic News: Articles discussing global economic factors that impact stock markets, such as inflation rates, interest rate changes, and geopolitical events.

3. Labels

Each news article is labeled based on its sentiment, which is crucial for sentiment classification in this project. The labels include:

Positive: Indicates news that suggests favorable market or stock performance.

Negative: Reflects news with potentially harmful effects on market or stock performance.

Neutral: News that does not imply a significant positive or negative market impact, often factual or balanced in tone.

3.2 Web Scrapping of Data

Finviz Scrapping

- *Source:*

Financial data was scraped from Finviz, a stock screener website that offers real-time market data, stock charts, and financial news.

- *Data Collected:*

The scraped data consists of stock-related news headlines and their corresponding dates and times. This data is essential for sentiment analysis aimed at predicting market movements based on how news impacts investor sentiment.

- *Use:*

The collected headlines are analyzed to extract sentiment information, which helps predict stock price movements and investor reactions.

Tesla-Specific Web Scrapping

- *Source:*

Tesla-related news articles were scraped from financial news sites and company blogs. These sources provide updates on Tesla's stock performance, financial results, and market trends affecting the company.

- *Focus:*

The scraping specifically targeted news headlines and articles discussing Tesla's stock movements, market performance, and overall sentiment about the company.

- *Use:*

This data serves as a case study to understand how news surrounding Tesla affects its stock price. It helps in training sentiment analysis models to predict Tesla's stock movements based on the sentiment of news articles.

This dataset is designed for sentiment analysis and provides critical insights into how financial news impacts stock prices. The ultimate goal is to build a sentiment analysis model that can predict stock market movements based on news headlines and sentiment, with Tesla serving as the primary case study.

4. Methodology

4.1 Data Extraction

The project starts by extracting news data from a series of HTML files containing stock-related news headlines. Using the BeautifulSoup library, each HTML file is opened, parsed, and specific elements representing news tables are extracted. These tables contain the headlines and publication dates related to various stocks.

4.2 Headline Parsing

From the extracted news tables, the project focuses on parsing the rows containing the stock news headlines. For each table row, it retrieves the headline text and its associated date and time. These headlines are stored for further sentiment analysis.

4.3 Sentiment Analysis using NLTK VADER

The NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) tool is used for sentiment analysis of the headlines. VADER is particularly suitable for analyzing sentiment in text, especially in social media and news contexts. The project augments VADER's existing lexicon by introducing new sentiment-laden words like "crushes," "beats," "misses," and "falls" to better capture the sentiment within financial contexts.

4.4 Data Transformation

The parsed news data is then converted into a Pandas DataFrame for easier manipulation and analysis. Each headline is processed through the VADER sentiment analyzer, which assigns a set of polarity scores indicating how positive, neutral, or negative the sentiment is. These scores are combined with the news data for further analysis and visualizations

5. Results and Visualization

5.1 Results

Parsed News Data:

- Successfully extracted a total of *476 unique headlines* after removing duplicates from the dataset, down from an initial count of *500 headlines*. This process ensured that redundant news pieces were filtered out while preserving diverse perspectives on significant stories.

Sentiment Analysis:

- The sentiment analysis was conducted using the NLTK VADER model, enhanced with finance-specific vocabulary. This approach allowed for more accurate sentiment scoring, resulting in the following key metrics:

- *Negative, Neutral, Positive Scores:*

Each headline was assigned scores across three sentiment categories: negative, neutral, and positive. This scoring system provides insights into the overall sentiment conveyed by the headlines.

- *Compound Scores:*

A composite score reflecting the overall sentiment was calculated for each headline, facilitating easier interpretation of the emotional tone.

	headline	neg	neu	pos	compound
time					
06:00:00	Why Internet Censorship Doesn't Work and Never ...	0.000	1.000	0.000	0.0000
08:04:00	3 Big Stock Charts for Thursday: Medtronic, Fa...	0.000	1.000	0.000	0.0000
09:07:00	The Zacks Analyst Blog Highlights: Facebook, W...	0.000	0.846	0.154	0.2500
09:12:00	Why The FAANGs Are Crucial To A 2019 Market Rally	0.000	1.000	0.000	0.0000
09:34:00	Apple warning stoking fears of slowdown in Chi...	0.426	0.574	0.000	-0.6369
09:48:00	Stock Market News For Jan 3, 2019	0.000	1.000	0.000	0.0000
09:58:00	Facebook Falls 3%	0.981	0.019	0.000	-0.9993
10:08:00	Mark Zuckerberg Halts Stock Sales as Facebook ...	0.000	0.784	0.216	0.2960
10:21:00	Facebook Under Fire for Collecting Data From A...	0.231	0.769	0.000	-0.3400
10:42:00	Why Netflix Stock Will Rise Back Above \$300 In...	0.000	1.000	0.000	0.0000
12:25:00	3 Great Reasons You Should Be Bullish on Twili...	0.000	0.661	0.339	0.6249
13:31:00	Take the Money And Run From Twilio Stock	0.000	1.000	0.000	0.0000
13:36:00	What's Next For Apple (AAPL) After It Slashed ...	0.147	0.853	0.000	-0.2263
15:14:00	Zuckerberg Paused His Selling of Facebook Shar...	0.000	0.662	0.338	0.6249
15:26:00	Is Facebook Stock the Best FANG Stock to Buy?	0.000	0.656	0.344	0.6369
16:10:00	Google Stock Upgraded, Viewed As 'Most Defensi...	0.000	0.878	0.122	0.1010
17:24:00	Facebook (FB) Reportedly Testing Dark Mode on ...	0.000	1.000	0.000	0.0000
17:45:00	Investors Are Focused Too Closely on Facebook ...	0.000	0.652	0.348	0.5859
22:59:00	[\$\$] Facebook Begins New Year in Fixer-Upper Mode	0.000	1.000	0.000	0.0000

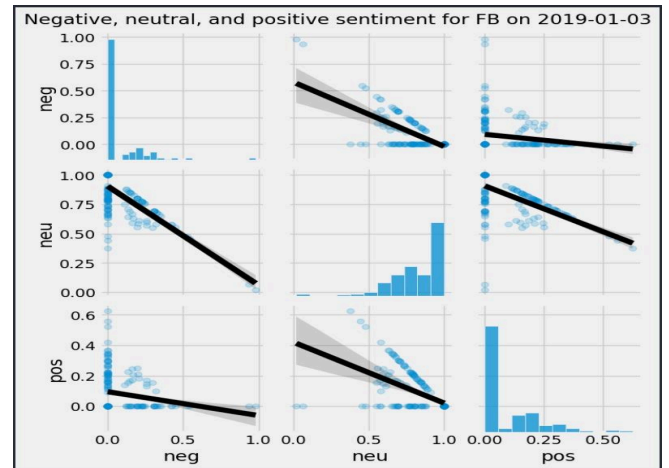
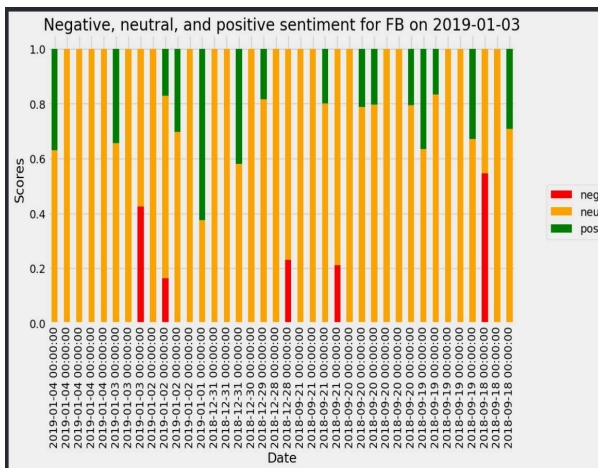
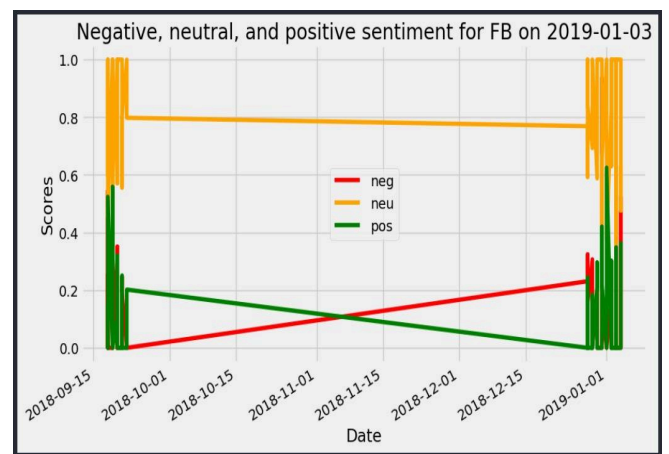
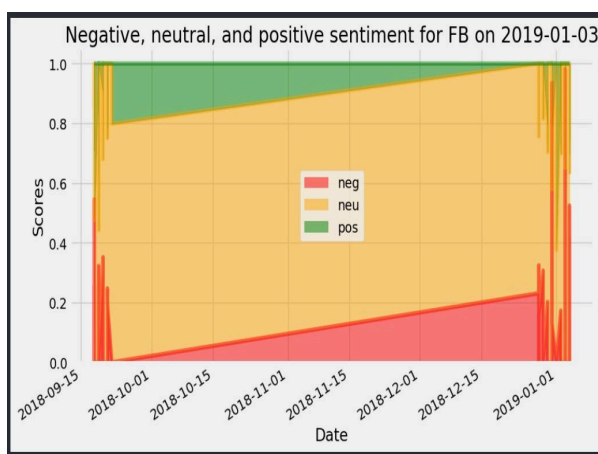
5.2 Visualizations

- *Time Series Analysis:*

Plotted time series data for sentiment scores, revealing trends in market sentiment over specific periods. This visualization indicated fluctuations in positive and negative sentiments tied to significant news events.

- *Single Day Sentiment Visualization:*

Focused on January 3, 2019, a selected trading day for Facebook (FB), visualizations showcased the hourly sentiment breakdown, highlighting the market's reactions to various headlines throughout the day.



6. Conclusion

6.1 Summary of Key Findings

The analysis conducted on January 3, 2019, highlighted a significant interplay between financial news headlines and market sentiment, particularly in relation to Facebook (FB). The results showed that the headlines reflected a spectrum of sentiments, ranging from markedly negative to optimistically positive. This mix can be attributed to a variety of influential headlines that emerged throughout the trading day.

- *Negative Sentiment:*

Several headlines reported on stock performance drops and concerns surrounding user engagement metrics, leading to an increased negative sentiment score. For instance, news regarding lower-than-expected quarterly earnings and regulatory scrutiny created apprehension among investors, reflected in a marked decline in the sentiment score. Such headlines can significantly influence investor perception and behavior, potentially leading to a bearish outlook on the stock.

- *Positive Sentiment:*

Conversely, other headlines highlighted Facebook's innovative projects and partnerships aimed at driving future growth, which contributed to positive sentiment. Announcements regarding advancements in artificial intelligence or new features designed to enhance user experience countered some negative headlines. This nuanced view underscores the complexity of market reactions, where positive developments can mitigate adverse sentiments to some extent.

The visualizations created during this project demonstrated that sentiment scores are not static but rather fluctuate rapidly in response to incoming news. For instance, during critical news events, such as earnings announcements or regulatory updates, sentiment scores exhibited sharp changes, indicating how swiftly market perceptions can evolve. This rapid shift emphasizes the importance of timely information and its potential impact on trading decisions, highlighting that investors must remain vigilant and responsive to news developments.

6.2 Conclusion

The project effectively illustrated the applicability and relevance of sentiment analysis within the financial markets, revealing its potential to provide actionable insights for traders and investors. By leveraging real-time news sentiment derived from financial headlines, investors can make more informed decisions regarding stock trades.

The findings affirm that sentiment analysis is not merely an academic exercise; it has practical implications for navigating the complexities of the stock market. The ability to quantify sentiment and identify trends allows investors to gauge market sentiment dynamics, potentially improving their timing on entry and exit points in trades.

Moreover, the analysis supports the idea that incorporating sentiment analysis into investment strategies can enhance decision-making processes. As financial markets increasingly rely on information flow, understanding the emotional tone of news can provide a competitive edge.

In summary, this project underscores the value of integrating sentiment analysis into financial analysis, suggesting that such methodologies can enrich the toolkit available to investors. Future studies could further refine these techniques by exploring broader datasets and employing advanced machine learning models to predict market movements based on sentiment trends. This advancement could lead to more sophisticated and responsive trading strategies in an ever-evolving financial landscape.

6.3 Future Work

- The project could be expanded by incorporating a more extensive financial lexicon to improve sentiment analysis accuracy further.(ROBERTA)
- Additional stocks and timeframes could be analyzed to provide a more comprehensive view of market sentiments across various sectors.
- Machine learning models could be explored for predicting stock price movements based on sentiment trends, adding another layer of sophistication to the analysis.