sciendo Journal of Social and Economic Statistics

Vol. 11, No. 1-2, Summer-Winter 2022
DOI: 10.2478/jses-2022-0003

OPEN ACCESS CC BY

*USING DATA MINING IN THE SENTIMENT ANALYSIS PROCESS ON THE FINANCIAL MARKET**

Marian Pompiliu CRISTESCU[a*], Raluca Andreea NERIȘANU[b], Dumitru Alexandru MARA[c]

[a) b) c)] Lucian Blaga University of Sibiu, Romania

Abstract

Sentiment analysis refers to the analysis of human opinions and sentiments that are expressed in written text, being also a part of the Natural Language Processing (NLP) tasks. Sentiment analysis can be applied in different domains, especially in the corporate marketing and sales, the healthcare system or the financial market analysis. In this paper we aim to highlight how data mining is able to extract the sentiment score from a financial platform that shows the major headlines regarding stocks, in order to highlight the publications' positive or negative opinion over a stock. In order to gain the sentiment score we have scraped text data from the platform Finviz from which the polarity of the opinion may be extracted. We have also used Valence Aware Dictionary for Sentiment Reasoning (VADER), by running a Python script using the BeautifulSoup library. After that we have used Pandas (Python Data Analysis Library) to analyse and obtain a sentiment score on the article headlines. Results show that the script is able to generate the sentiment score for various selected stocks, while also showing graphical diagrams for the past and future trend of the stock, in terms of overall opinion on the stock performance.

Keywords: VADER, sentiment analysis, data mining, BeautifulSoup library, Finviz, Pandas
JEL Classification: C63Computational Techniques • Simulation Modeling

* Corresponding author, **Dumitru Alexandru Mara** – dumitrualexandru.mara@ulbsibiu.ro

*This paper was presented at the International Conference on Applied Statistics ICAS 2022. The authors thank participants for their useful feedback.

1. Introduction

The financial crisis of 2008, for instance, showed that the creation of risk models with dynamics for abnormal circumstances does not ensure that financial institutions will be successful in controlling extremely high systematic risks. The fact that risk managers still solely take stock data into account when choosing their target portfolio can be one of the arguments in favour of this view (Fan and Gu, 2003). They frequently fail to consider the risks associated with assets that are not part of the portfolio; as a result, the size of the portfolio for risk analysis may be rather minimal. It's possible that the so-called small portfolio risk analysis fails to adequately represent portfolio risk dynamics, particularly systematic risks that come with the market.

Therefore, risk managers can take into account latent risks that cannot be contained in the portfolios of interest by using data science and big data techniques in the financial industry.

The amount of textual data on the Internet is expanding rapidly, and many firms and organizations are striving to leverage this data stream to extract people's opinions on their goods (Sheela, 2016).

In (Provost and Fawcett, 2013) are provided examples of uses of data mining, including targeted marketing, internet advertising, and cross-selling recommendations. In addition, data mining is used for overall customer relationship management to monitor customer behaviour to minimize customer churn and optimize estimated customer value. Data mining is used by the financial industry for credit assessment and trading, as well as for fraud detection and workforce management. From marketing to supply chain management, major retailers like Walmart and Amazon are using data mining in their operations. Numerous companies have intentionally differentiated themselves with data science, often to the point of becoming data mining corporations.

Therefore, in this paper, we aim to achieve both numerical and visual results of the sentiments that are expressed through financial news regarding specific stocks. To do this, we intend to use the FinViz platform to subtract the sentiment from the headlines of the financial news, in association with a specific stock and thus generate both numerical values of the sentiment in consecutive days and visual graphics of it, in the form of graphs, in order to foster a clear view of the sentiment trendline.

2. Data mining

Data mining is defined by (Ertel, 2017) as “the task of a learning machine to extract knowledge from training data” (p.179). Ertel also defined data mining as the process of acquiring information from data using statistics or machine learning in the setting of massive amounts of data at an affordable price (Ertel, 2017). In order to distinguish from statistics to big data, the dimensions of the concept of big data refer to: (1) Volume (the size of the data), I

terms of exceeding the limit that can be managed by a conventional database software (Manyika *et al.*, 2011; Kabir and Carayannis, 2013); (2) Variety of the data, considering a variety of data forms (numerical scaled data, scaled or non-scaled data), data sources (internal or external), data formats (photographs, videos, text or sounds) (Kabir and Carayannis, 2013; Sangeetha and Sreeja, 2015; Russom, 2021), content format (semi-structured, unstructured, or structured); data sources (tweets, blogs, product assessment, and social network data) (Assunção *et al.*, 2015); (3) Velocity (rate and cadence of data reception), such as human actions, machine outputs, web and social media locations (Sangeetha and Sreeja, 2015); (4) Veracity (the dependability of the data) (Sangeetha and Sreeja, 2015; Barham, 2017); (5) Value of the gathered data, (Barham, 2017; Russom, 2021).

In recent years and today, the amount of data created has increased dramatically. This large amount of data may be gathered from several sources such as Web and social media, Machine to Machine, Big Transaction data, Biometrics and Human generated data.

Web and social Media data can include clickstream data, social media platform postings, content of websites and many more. Machine to Machine can include, utility smart meter readings, radio frequency identification readings, GPS signals and other sensor readings. Big Transaction data may include telecommunications call detail record, healthcare claims, utility billing records. Data can also be generated by humans through their voice recordings, email, SMS texts, electronic medical records and other (Mohanty, Senapati and Lenka, 2013; Shim *et al.*, 2015).

There are other techniques to extract data from websites, but web scraping has shown to be the most effective. Using programs known as crawlers, this method is utilized to extract data from websites. One of the benefits of Web scraping is that once the code is created and executed, it is possible to automatically extract data from the defined domains (websites) (Prathi, Raparthi and Gopalachari, 2020).

There are multiple techniques for implementing web scraping. However, despite the fact that this study will not focus on them, it is essential to note that mastering this technique is crucial to the success of business decisions.

3. Sentiment analysis

Sentiment Analysis (SA) or Opinion Mining is a set of techniques used to analyse opinionated text containing people's opinions about various entities, including products, services, organizations, and individuals, among others (Gandomi and Haider, 2015).

Various concepts associated with sentiment analysis have been uncovered through a review of the literature. Text analytics, consumer analytics, and data mining are included.

Text analytics (TA) is the collection of information from textual data generated by users, such as product reviews, social network posts, online forums, emails, blogs, survey responses, and reports, among others. In addition, TA enables organizations to effectively handle and

manage large volumes of human-generated content that have the potential to become valuable insights and information for the organization (Gandomi and Haider, 2015).

Consumer analytics can be defined as the extraction of unperceived consumer insights from the complexity of Big Data and the exploration of these insights via a profitable interpretation (Erevelles, Fukawa and Swayne, 2016).

Sentiment analysis can be applied in a variety of contexts, including the commercial product industry, politics (citizens' opinions on certain topics and political elections, among others), and stock market forecasting (Hemmatian and Sohrabi, 2019). This study focuses on showing how data mining can extract the sentiment score from a financial platform that displays the most prominent headlines about stocks in order to emphasize the publications' favourable or negative view of a stock.

Among the first techniques created for identifying opinions in a text are solutions based on the extraction of linguistic ideas, which typically exploit the presence in the text of words with sentiment value kept in sentiment dictionaries. Godbole, Srinivasah, and Skiena describe the building of this dictionary (Godbole, Manjunath and Skiena, 2007). The method extends a collection of sentiment-valued keywords by examining the synonymy and antonymy associations produced by the WordNet lexicon with respect to the keywords (Miller, 1995). In addition, it is established that the emotion meaning of each identified word is inversely related to its reported distance from the keywords. Taking into consideration the tree relationship unique to the investigated problem, the emotion polarity of a word may change many times up to a certain distance. To prevent contaminating the sentiment lexicon, only sentiment words with a restricted number of observed modifications at the level of polarity are retained.

Prabowo and Thelwall describe a set of sentiment classifiers that split papers into Positive and Negative classes based on the frequency of occurrence of sentiment words in close proximity to other terms in the document's composition (Thelwall, Homsi and Prabowo, 2009). The suggested classifiers utilize either observed frequencies related to terms in a sentiment lexicon or frequencies given by Google or Yahoo search engines relative to a limited selection of sentiment-laden phrases. Finally, the polarity of the text is determined by the sentiment class linked with the majority of the document's terms.

Following the development of language methods mostly based on unsupervised approaches, models based on machine learning often address the challenge of recognizing sentiment polarity in a supervised manner. Ma, Yuan, and Wu give an unsupervised example for the machine learning problem through a multi-dataset investigation (Ma *et al.*, 2017). The end objective of this study is to examine, from the standpoint of document-level sentiment analysis, various bag-of-words input formats and clustering techniques. Bag-of-words representations regard documents as a collection of words whose order is unimportant. The authors conclude, based on the datasets analysed, that the Kmeans model and its version RB-Kmeans (Michael Steinbach, George Karypis and Vipin Kumar, 2000) produce the best results,

coupled with the representations derived using DPH (Divergence from randomness) scores (Giambattista, Amati Giuseppe *et al.*, 2008).

In terms of supervised learning, the most prevalent approaches are the naïve Bayesian classifier or Support Vector Machine (SVM) models (Cortes and Vapnik, 1995), maximum Entropy (Berger, Della Pietra and Della Pietra, 1996) and logistic regression. Read proposes a research in which SVM and naïve Bayesian classifiers are used across a collection of one-dimensional representations that indicate the binary existence of vocabulary terms in the input texts (Cordeiro *et al.*, 2014). According to Denecke (2008), input representations are derived using the SentiWordNet sentiment lexicon (Esuli and Sebastiani, 2006) word-level sentiment scores (Denecke, 2008). On the basis of this information, a logistic regression is then conducted to construct a set of probabilities that indicate the input's membership in the studied sentiment classes.

In addition to the two classifiers employed by Gautam and Yadav examine the Maximum Entropy model (Gautam and Yadav, 2014). Moreover, the authors demonstrate that semantic analysis based on the similarities between words is superior to the three models in terms of extracting sentiment polarity from tweets. In the training phase of semantic analysis, it is assumed that if an adjective appears in a text with a particular sentiment orientation, then the adjective must have the same orientation. The selection of adjectives is based on their descriptive capacity and their frequently emotive meaning. Later, during the test phase, an input text is assigned a certain sentiment polarity based on the similarities between its adjectives and those detected during the training phase.

The performance of machine learning-based sentiment classifiers is frequently dependent on how the input is described (Balazs and Velásquez, 2016). Neural networks are believed to be able to overcome this constraint, while belonging to the field of machine learning, by automatically extracting usable information from the input without the need for its prior preparation. This allows neural networks to generalize more effectively than other machine learning models (Chaturvedi *et al.*, 2018).

As most linguistic approaches are based on syntactic constructs that combine aspects with the remainder of the text, the majority of methods in this category are unsupervised. The most typical strategy in linguistic approaches consists of selecting and filtering potential terms in two rounds to extract explicit features. Yi, Nasukawa, Bunescu, and Niblack offer one of the earliest approaches used to detect explicitly specified components in the text based on the typical two-step methodology (Yi *et al.*, 2003). Yi et al. pick potential aspects on the basis of this concept, which is based on the observation that aspects are predominantly represented by nouns and is heavily utilized in linguistic approaches. The potential elements are then screened based on their degree of relevance to the topic of study.

The approach described by Hu and Liu is intended to extract not just the often-noticed aspects of the text, but also the less frequently observed aspects (Hu and Liu, 2004). For

frequent feature identification, Liu, Hsu, and Ma association rules are utilized (Liu *et al.*, 1998). Then, features that do not consistently occur in the same sequence or are incorporated in other aspects are eliminated. Hu and Liu define adjectives (which are supposed to be opinion-bearing) and consider the nouns associated to them to be the sought-after aspects in terms of extracting less common features (Hu and Liu, 2004).

BERT is widely used as the initial layer of neurons necessary to build new vector representations for each word in a sentence (Devlin *et al.*, 2019; Colasanto *et al.*, 2022). Due to the fact that BERT is a model constructed only on the attention mechanism, the same word might have distinct representations depending on its context. The BERT encoder is also utilized in the approaches reported in (Phan and Ogunbona, 2020; Karimi, Rossi and Prati, 2021). In the instance of Phan and Ogunbona's technique, the layer with output neurons where the SoftMax function is applied runs on a set of representations generated by RoBERTa, an upgraded form of the BERT model proposed by (Liu *et al.*, 2019). In addition to the RoBERTa representations, the model employs a second set of learnt representations to hold information on POS tags and word dependency relationships. In the method presented by Karimi et al., Li et allinear's approach is placed within an adversarial training model (Karimi, Rossi and Prati, 2021). Specifically, the cost function generated by applying the BERT coder to a layer of output neurons whose activation function is SoftMax is utilized to generate the perturbation terms required to generate a new fake input. The BERT model and output layer are reapplied to the new input, yielding a new cost function that sums with the cost function derived from the real input. The new cost function is applied to all model weights to alter them. In conclusion, the model provided by Karimi et al. is utilized not only for the prediction of aspects, but also for the identification of emotional polarity.

Xu, Liu, Shu, and Yu suggest substituting the second objective with a classification of the text from the perspective of the domain of belonging, given that the BERT model is trained to anticipate missing words in a text and to detect whether two sentences are sequential inside the same text (Xu *et al.*, 2021). DomBERT is a novel model presented exclusively for finding aspects using a standard technique, similar to the solution offered in (H. Xu *et al.*, 2020)

Similar to the concept utilized in (C. Xu *et al.*, 2020), the technique presented in (Wu *et al.*, 2019) is based on estimating the probability that a word denotes the beginning or end of an aspect utilizing two layers of output neurons operating on a collection of BERT-like vector representations. Since a sentence might have a variable number of aspects, Wu et al. select just the shortest and non-overlapping aspects (Wu *et al.*, 2019).

In light of the efficiency of the basic architecture suggested in (Li *et al.*, 2019), a method for feature extraction beginning with the BERT encoder is proposed in the current study. In the present literature, there is a propensity to discover characteristics utilizing language techniques or machine learning-based techniques. The suggested model is defined in such a manner as to take advantage of both the syntactic and contextual information discovered by the attention

processes of the BERT model as well as linguistic ideas. While the BERT model gathers relevant information, language concepts are utilized as validators to ensure that features are correctly identified. Instead of allowing a neural network learn the relevant information on its own, a preferable method would be to guide it based on information particular to the situation being examined so that it can more quickly recognize the target information. On the basis of the described architecture, the suggested model may be compared to standard linguistic methodologies studied in two phases. In the present instance, the linguistic ideas continue to play the function of picking potential aspects, while the BERT model serves as a filtering strategy for identifying the final aspects.

Similar to the discovery of document-level sentiment polarity, linguistic methods were among the first approaches established for aspect-level sentiment categorization. The technique presented by (Hu and Liu, 2004) is comparable to the document-level sentiment analysis method introduced in (Thelwall, Homsi and Prabowo, 2009). Using the antonymy and synonymy relations specified in the WordNet dictionary, the approach determines the sentiment polarity for each adjective in an input instance, given a collection of keywords for which the sentiment orientations are known. The entire input is then assigned the polarity associated with the most common sentiment class detected at the adjective level. Despite the fact that the technique is aware of the aspects, their sense polarity remains consistent over the full input. When the number of positive adjectives equals the number of negative adjectives, only then can distinct polarities be assigned to different features of an input. In this scenario, the majority class of the linked adjectives determines the emotional polarity of a specific feature.

Lexicon-based approach is a strategy that assigns particular weights to each word in the text based on the polarity to which it belongs (negative, positive or neutral). Utilizing resources like as SentiWordNet or VADER (Valence Aware Dictionary for Sentiment Reasoning), among others, this process may be implemented (Hutto, C.J. and Gilbert, 2014; Al-Shabi, 2020).

4. Methodology

To obtain the sentiment score, we gathered text data from the Finviz, presented in (Figure no. 1) platform from which the opinion polarity may be inferred. In addition, we have utilized the Valence Aware Dictionary for Sentiment Reasoning (VADER) by executing a Python script utilizing the BeautifulSoup package. After analysing the article headlines using Pandas (Python Data Analysis Library), we obtained a sentiment score.

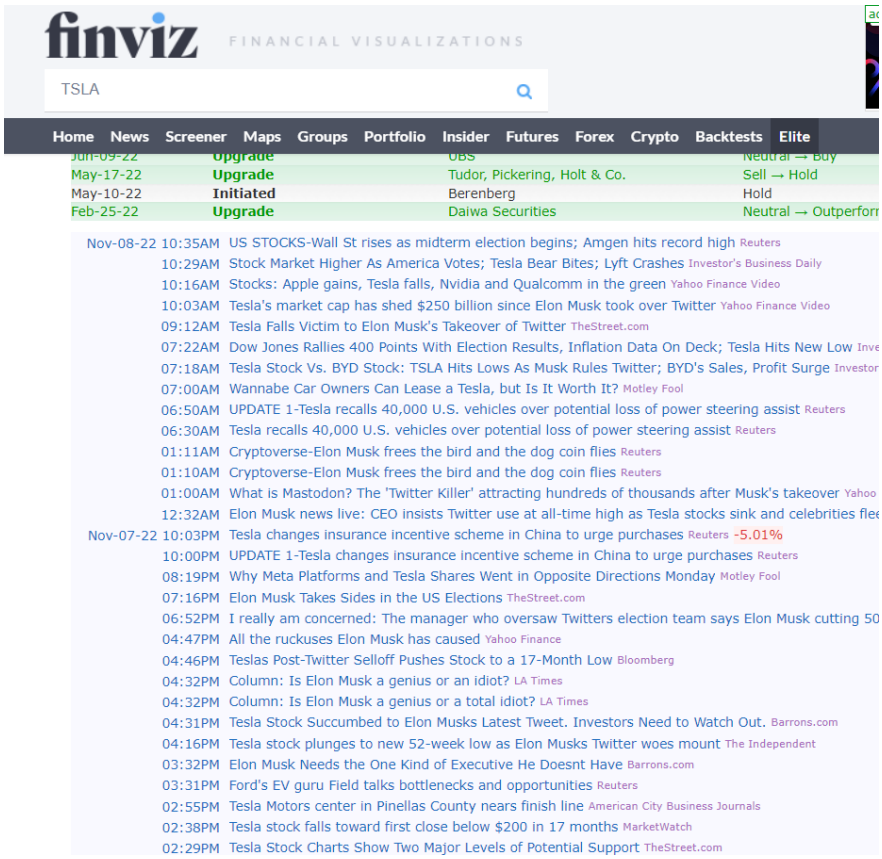


Figure no. 1: The Finviz Platform

Source: authors' computation

BeautifulSoup can interpret any input by employing simple Pythonic idioms and methods to construct a searchable and navigable parse tree, as (Figure no. 2) shows.

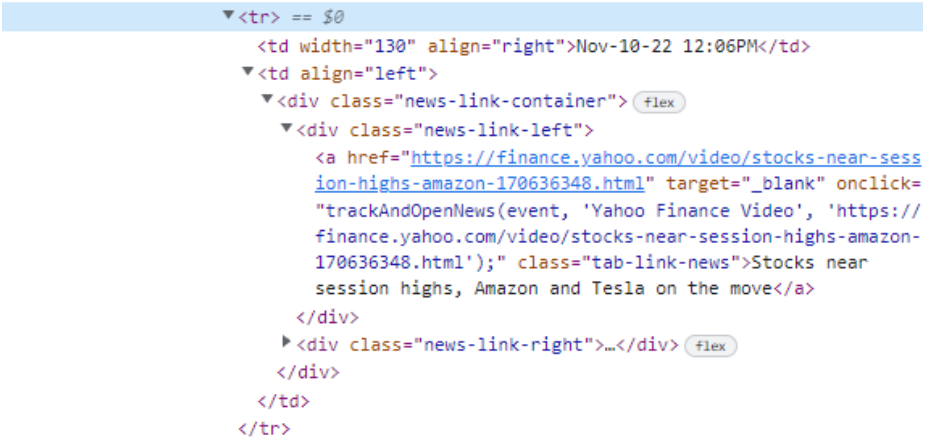


Figure no. 2: The HTML code of a headline shown on Finviz

Source: authors' computation

First of all we imported the required libraries that we will use and defined a variable with the url that refers to the Finviz platform. As we can see in, the platform allows the input of a ticker withing the GET variable "t". In (Figure no. 3) that one is "TSLA".

```
from statistics import mean
from urllib.request import urlopen, Request
from bs4 import BeautifulSoup
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import pandas as pd
import matplotlib.pyplot as plt

finviz_url = 'https://finviz.com/quote.ashx?t='
tickers = ['TSLA']

news_tables = {}
for ticker in tickers:
    url = finviz_url + ticker

    req = Request(url=url, headers={'user-agent': 'my-app'})
    response = urlopen(req)
```

Figure no. 3: Importing the required libraries and defining the source of our data

Source: authors' computation

As it can be seen in (Figure no. 4), we parsed into an array the table rows of the table that Finviz outputs.

```
    html = BeautifulSoup(response, features='html.parser')
    news_table = html.find(id='news-table')
    news_tables[ticker] = news_table
parsed_data = []

for ticker, news_table in news_tables.items():

    for row in news_table.findAll('tr'):

        title = row.a.text

        date_data = row.td.text.split(' ')

        if len(date_data) == 1:
            time = date_data[0]
        else:
            date = date_data[0]
            time = date_data[1]
        print (date+" | "+title)
        parsed_data.append([ticker, date, time, title])
```

Figure no. 4: The parsing of headlines from Finviz into an array using Python

Source: authors' computation

A model is required for sentiment analysis. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model based on rules that can be used for general sentiment analysis. Its sensitivity regards polarity and the emotion's strength can be used for unlabelled text data. VADER is included in the NLTK package which is a platform for building Python programs that facilitate working with human language data (https://www.nltk.org/). VADER is different from LIWC though better generalization for different domains and becoming more attentive to expressions of emotion in social media contexts. Hutto and Gilbert (2014) were able to build and empirically validate a list of lexical features that are uniquely sensitive to sentiment in microblog-like contexts. Thus, VADER can be utilized for the sentiment analysis of financial news headlines published online and shared on social media. It is crucial to note, however, that on certain dates, some stocks were not mentioned by the major financial news outlets from which Finviz obtains its data, and the sentiment score was therefore assumed to be zero.

After the parsing of the headlines, we analyse the text using the VADER model, as can be seen in (Figure no. 5).

```
df = pd.DataFrame(parsed_data, columns=['ticker', 'date', 'time', 'title'])

vader = SentimentIntensityAnalyzer()

f = lambda title: vader.polarity_scores(title)['compound']

df['compound'] = df['title'].apply(f)

df['date'] = pd.to_datetime(df.date).dt.date

plt.figure(figsize=(10,8))
mean_df = df.groupby(['ticker', 'date']).mean().unstack()
mean_df = mean_df.xs('compound', axis="columns")
print(mean_df)
mean_df.plot(kind='bar')
plt.show()
```

Figure no. 5: The analysis of the headlines using VADER

Source: authors' computation

In order to summarise the used procedure, we have introduced the diagram from (Figure no. 6).

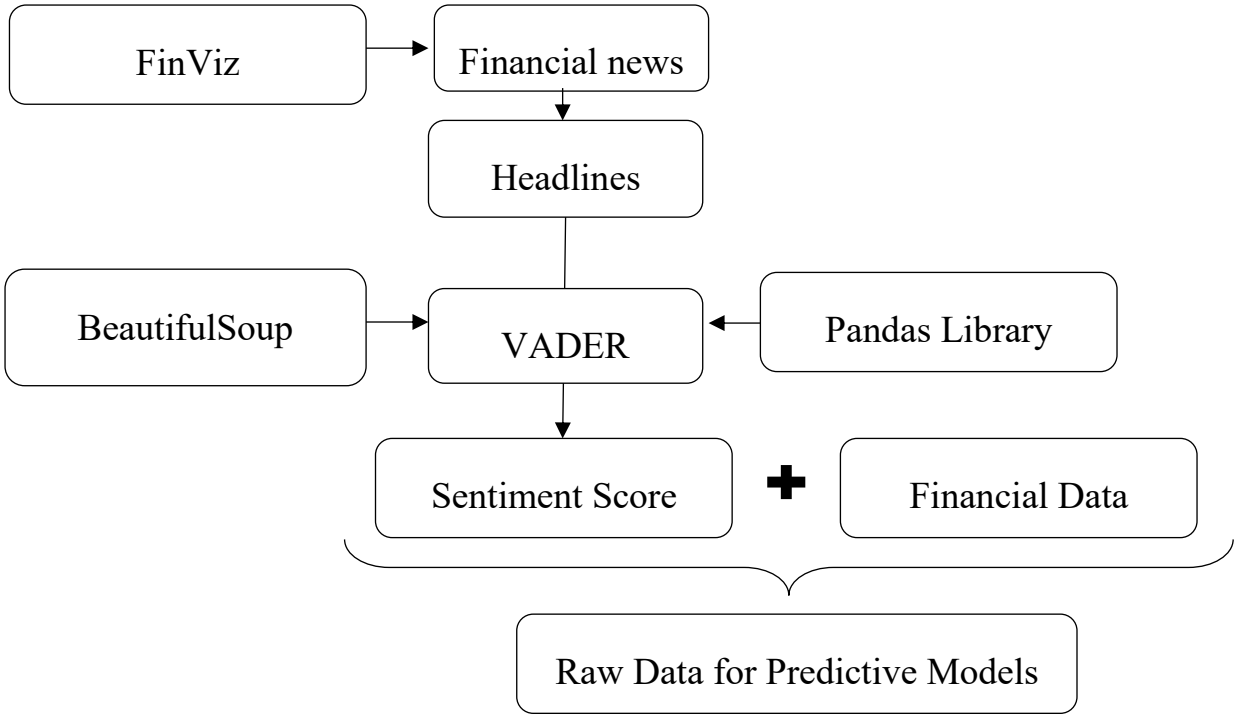


Figure no. 6: Research design

Source: authors' computation

After gaining the raw data for predictive models, both parametric and nonparametric models may be used to forecast future stock behavior (Giudici, Mezzetti and Muliere, 2003).

5. Results

The Python script is able to output the analyzed text and calculate the sentiment score for the input.

Also, the script is able to generate and show a graph for the recent days in which a stock has been covered in financial news highlighting the evolution of the sentiment score of the headlines regarding the stock. This can provide a fast way for a user to see graphically the recent trend in the financial news publications regarding a specific stock. Of course, this script can be automated further in order to export the data in other formats.

These sentiment scores were calculated by the script and provided both numerically and shown graphically, as shown in (Figure no.7), that can furthermore be processed and for example correlated with the stock opening and closing prices in the stock market. Therefore, from Figure no.7 we can observe that a positive sentiment score was achieved in 03.11.2022

and 08.11.2022 as titles in 03.11.2022 included phrases like "Why (…) Tesla (…) Stocks Were Up This Morning" and "Elon Musk Revamps Twitter with help from Tesla Staff". In 04.11.2022 the sentiment was neutral, with both positive and negative headlines, in 05.11.2022, 06.11.2022 and 07.11.2022 there was registered a high negative sentiment score, as titles included phrases like "Tesla stock plunges" and "Tesla stock falls".

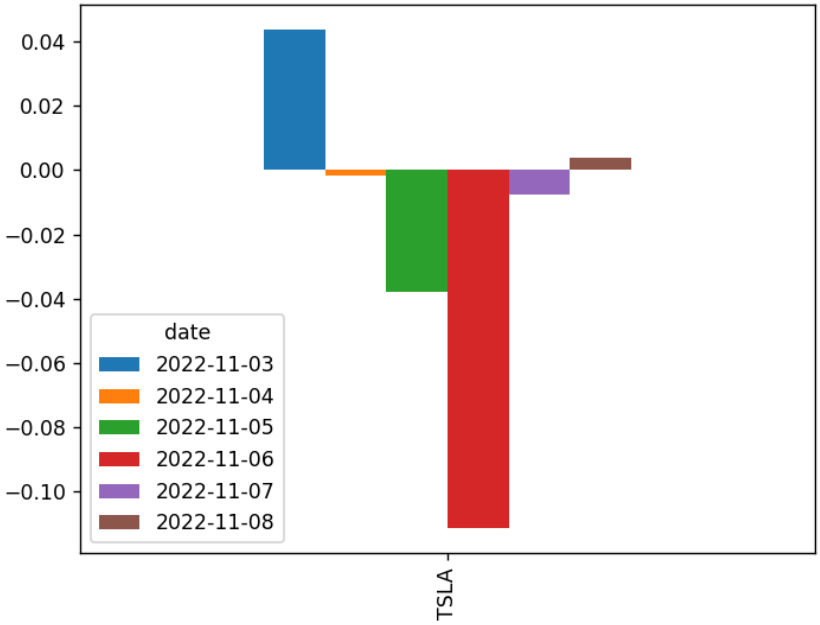


Figure no. 7: Tesla stock sentiment graph for the period 03-08.11.2022

Source: authors' computation

After extracting the text from Finviz for each selected firm and analysing the sentiment of the articles between 09.08.2022-24.08.2022, the numerical values were imported in statistical software tools, that generated the line graphs presented below. Thus, for all of the analysed stocks, we obtained an average sentiment score of 0.04 and an average volatility score of 0.19 for the news. As can be observed in (Figure no. 8), the mean sentiment score for F (Ford Motor Company) stock was 0,12 suggesting a positive perception of it by the financial news, having the lowest volatility among the other stocks as (Table no. 1) shows.

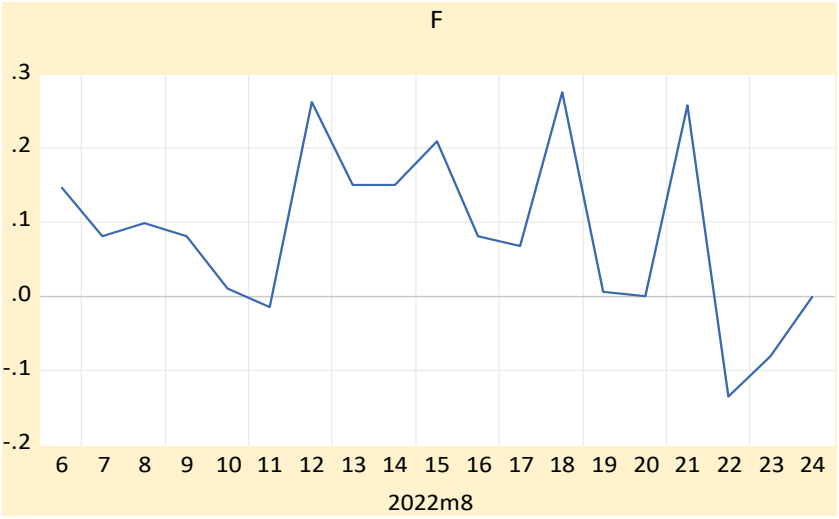


Figure no. 8: Sentiment score for F stock

Source: authors' computation

The mean sentiment score for AMD (Advanced Micro Devices) stock had a value of 0,12 and a volatility of 0,23 (value rounded from 0,231), the highest one among the other stocks (Figure no. 9).

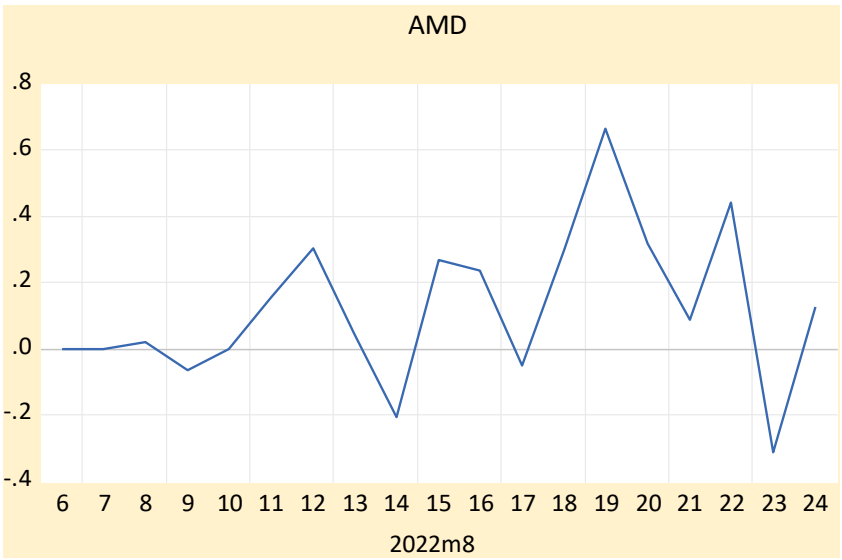


Figure no. 9: Sentiment score for AMD stock

Source: authors' computation

The mean sentiment score for NFLX (Netflix) stock was 0,01 suggesting a positive news sentiment, presenting a volatility of 0,18 during the analysed period which is below the average volatility of 0,19 (Figure no. 10).

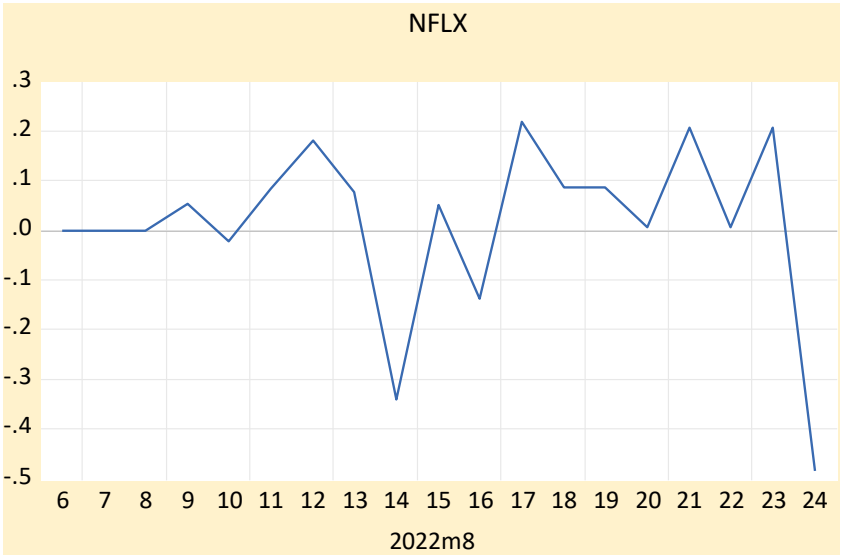


Figure no. 10: Sentiment score for NFLX stock

Source: authors' computation

The mean sentiment score for NVDA (Nvidia) stock was 0,04 suggesting a positive sentiment of the news related to this company's stock and a volatility of 0,16, a relatively low one (Figure no. 11).

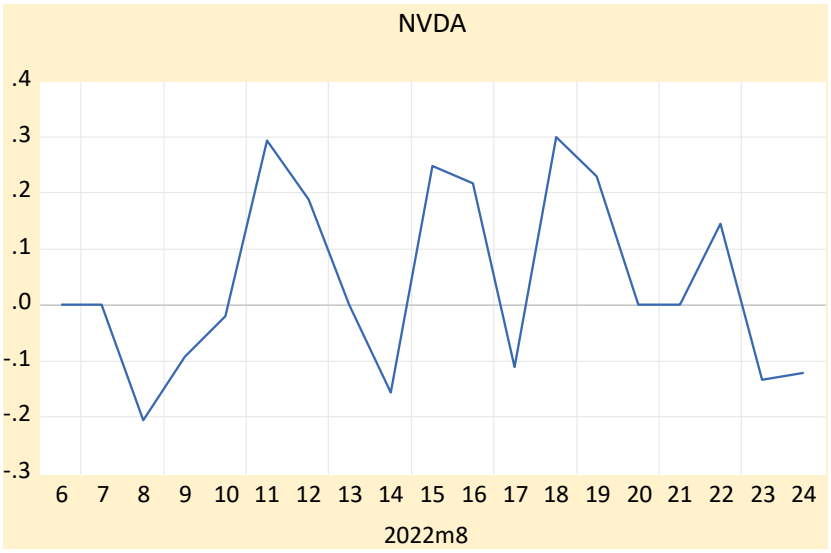


Figure no. 11: Sentiment score for NVDA stock

Source: authors' computation

The mean sentiment scores for GM (General Motors) stocks was -0,01, suggesting a negative perception of them. The news' sentiment score presents a volatility of 0,21 (Figure no. 12).

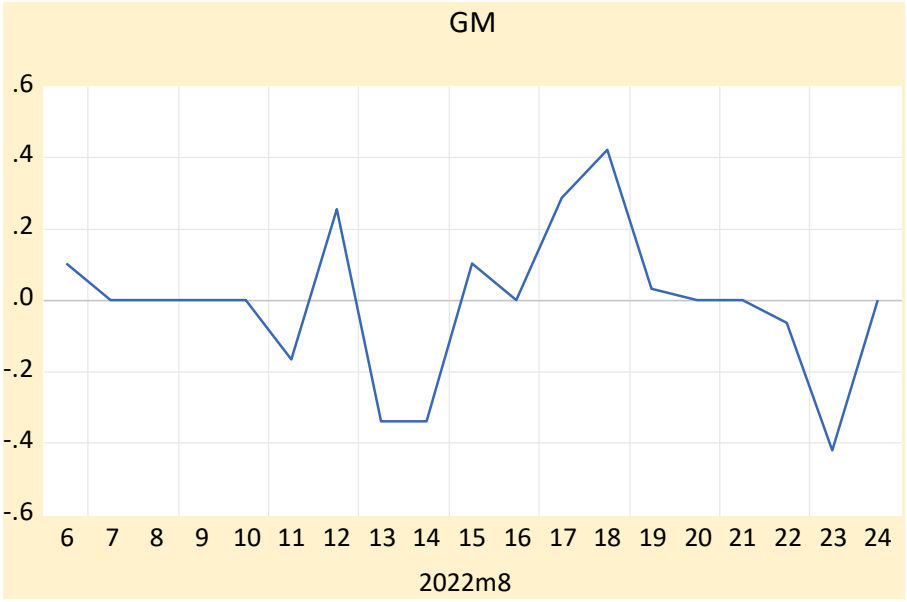


Figure no. 12: Sentiment score for GM stock

Source: authors' computation

On average the sentiment score of the news related to BRK-B (Berkshire Hatha-way Inc Class B) stock was 0,08 suggesting a positive sentiment and a volatility of 0,15, a relatively low one, below the average volatility of 0,19 (Figure no. 13).

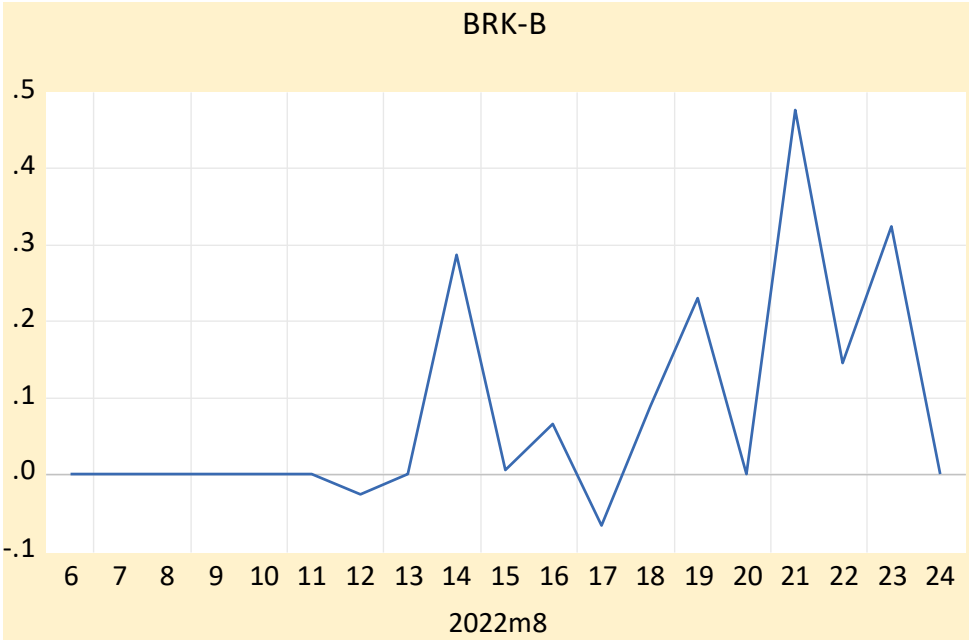


Figure no. 13: Sentiment score for BRK-B stock

Source: authors' computation

In (Table no. 1), most positive and negative averages of the stock volatility are exposed, along with the highest and lowest volatility for the analysed period.

Table no. 1. Volatility indicators for the analysed stocks

Indicator	Value	Stock
Most negative average	-0.006951	GM
Most positive average	0.122141	AMD
Highest volatility	0.23077957	AMD
Highest annualized volatility	3.663512101	AMD
Lowest volatility	0.115154285	F
Lowest annualized volatility	1.828017599	F

Source: authors' computation

While comparing the six line graphs above (Figure no. 8-13), some general conclusions can be drawn: (a) There was a neutral trendline in sentiment score in the first period that was analysed (06.08.2022-11.08.2022), except for NVDA; (b) The majority of the stocks had their peaks in sentiment score between 17.08.2022 and 19.08.2022, except for BRK, which had its positive peak on 22.08.2022, while F and GM registered negative sentiment scores in 22.08.2022; (c) The most volatile stock in terms of financial sentiment score was registered to be AMD, which ranged from -0.31 to 0.66, with the most positive value registered in 19.08.2022 and the most negative one in 23.08.2022, while also having the highest average sentiment score of the analgised stocks; (d) The less volatile stock was F, which ranged from -0.13 to 0.27 in terms of sentiment score, in the analysed period.

6. Discussion

We have used in this research the Beautiful Soup library in relation with the FinViz platform to gather the headlines of financial news on the selected companies, and afterwards we used VADER model within a Python Script to show the general sentiment regarding the events that may occur regarding the equities under analysis. In the end, we used Pandas to analyse and run a sentiment analysis on the article headlines.

Future stock trend analysis is a difficult task due to the numerous variables involved. We have hypothesized that news items and share prices are correlated and that the news may correspond with share price swings.

The data indicate that the sentiment scores varied dramatically from day to day. The average sentiment of market news between 06.08.2022 and 24.08.2022 was 0.047416, indicating that the news generated a positive attitude. The lowest sentiment score was -0,765000 for Sony Corporation in 24.08.2022 and the highest one for Advanced Micro Devices (AMD) with a sentiment score of 0,663850 in 19.08.2022. Also, these companies have the lowest and highest

average of the sentiment score, Sony Corporation having a -0,088537 score and AMD a 0,122141 score during the analysed period.

Afterwards, we determined that AMD stock had the most volatile sentiment score, while Ford Motor Company stock had the least volatile sentiment score. This can contribute to the notion that Ford Motor Company stock may be the safest investment in terms of volatility, given that the sentiment of its news headlines is quite consistent and the opinions of prominent financial news publications are not significantly divided.

Similar studies constructed sentiment indexes based on financial news gained from different markets (Wei *et al.*, 2017).

In order to predict the future behavior of the stock prices and to consider forecasting a profitable behavior of the investor, in (Theodorou *et al.*, 2021) additional signs of the stock behavior, in relation with the daily sentiment analysis were obtained.

In other studies, different stock indexes were used in order to forecast the future behavior of the stocks, by analyzing also the society opinion on the analyzed stocks, as for example in (Yıldırım, Toroslu and Fiore, 2021), where smart decision logic was used in order to incorporate "long short-term memories" into an up to five days predictive model. In contrast, in (Giudici, Mezzetti and Muliere, 2003) it was found that "a polarized" sentiment was important into determining speculative changes in the market than the full news volume. Similar, strong correlation was found between price volatility and sentiment disagreement (Siganos, Vagenas-Nanos and Verwijmeren, 2017).

Although our study used FinViz in order to gain financial news, other studies have used different platforms, such as Google or Wikipedia, showing that the number of the used databases is directly corelated with the model precision (Weng, Ahmed and Megahed, 2017). Also, other studies imply tensor decomposition and coupling matrices (Zhang et al., 2018).

7. Conclusions

The script we developed using free solutions manages to extract data from the web and inform the user regarding the recent sentiment of the financial news regarding a specific stock and can be adapted for other uses.

Sentiment analysis enables learning about the opinions of an audience regarding a product or service.

In combination with web scraping, can be used to generate the sentiment with which major financial publications inform their audience, being a tool that can be considered when forecasting future directions of the stock prices and making investment decisions.

One of the main advantages of sentiment analysis in financial decision making is that it can provide a more comprehensive and nuanced understanding of market sentiment than traditional financial metrics alone. For example, by analyzing large volumes of social media data, sentiment analysis can reveal patterns and trends in consumer sentiment that may not be apparent from traditional market indicators such as stock prices or trading volume. This can be particularly useful for identifying early warning signs of market shifts or for identifying new investment opportunities.

The main advantage of using the method presented in the paper is related to the quick return of the current general opinion on some object, such as stocks and the opportunity to use the society's opinion to reach some speculative profits. Also, the script provides useful visualization of the sentiment scores, and a predictable trendline.

However, sentiment analysis also has its limitations and drawbacks. One major disadvantage is that it can be difficult to accurately interpret the meaning of natural language text, particularly when the text is written in colloquial or informal language. Additionally, sentiment analysis can be prone to errors and biases, particularly when the data used to train the analysis algorithm is not representative of the population of interest.

Another disadvantage is that sentiment analysis can be easily manipulated. For example, if a company or individual wants to artificially inflate or deflate sentiment about a particular stock or market trend, they can do so by creating fake social media accounts or by posting fake news articles. This can lead to inaccurate or misleading sentiment analysis results.

Regarding the presented script, it does provide scores only for a short period of time, because the platform FinViz only returns a limited amount of news for a specific stock and for a limited number of days, returning only the most recent ones. That would be insufficient for the long-term analysis of sentiment regarding stocks. Also, the script must be run manually at a specific point in time in order to return recent results. This may present an opportunity to automate its execution at frequent intervals in order to collect and provide the sentiment scores for longer periods of time.

Also, future research and improvements one may focus on modifying the script to calculate the sentiment scores using data from social media platforms, such as Twitter. Moreover, future research may concentrate on diversifying the types of data sources in order to improve the accuracy of market forecasting, or it may employ a combination of methodologies to achieve improved forecasting.

References

Al-Shabi, M. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining, *IJCSNS International Journal of Computer Science and Network Security*, 20(1), January 2020.

Assunção, M. D. *et al.* (2015). Big Data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, 79–80, pp.3–15. doi: 10.1016/j.jpdc.2014.08.003.

Balazs, J. A. and Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey, *Information Fusion*, 27, pp. 95–110. doi: 10.1016/j.inffus.2015.06.002.

Barham, H. (2017). Achieving competitive advantage through big data: A literature review, *PICMET 2017 - Portland International Conference on Management of Engineering and Technology: Technology Management for the Interconnected World, Proceedings*, 2017-Janua, p. 1–7. doi: 10.23919/PICMET.2017.8125459.

Berger, A. L., Della Pietra, S. A. and Della Pietra, V. J. (1996). A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistice,* Cambridge, MA: MIT Press, 22(1), p. 39–71. Available at: https://aclanthology.org/J96-1002.

Chaturvedi, I. *et al.* (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, *Information Fusion*, 44, p. 65–77. doi: 10.1016/j.inffus.2017.12.006.

Colasanto, F. *et al.* (2022). BERT's sentiment score for portfolio optimization: a fine-tuned view in Black and Litterman model, *Neural Computing and Applications*. Springer London, 1. doi: 10.1007/s00521-022-07403-1.

Cordeiro, E. R. *et al.* (2014). Posttherapy Follow-up and First Intervention, *Prostate Cancer: Diagnosis and Clinical Management*, (June), pp. 211–229. doi: 10.1002/9781118347379.ch11.

Cortes, C. and Vapnik, V. (1995). 'Support-vector networks', *Machine Learning*, 20(3), p. 273–297. doi: 10.1007/BF00994018.

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis, in *2008 IEEE 24th International Conference on Data Engineering Workshop*. IEEE, pp. 507–512. doi: 10.1109/ICDEW.2008.4498370.

Devlin, J. *et al.* (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), pp. 4171–4186.

Erevelles, S., Fukawa, N. and Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing, *Journal of Business Research*, 69(2), p. 897–904. doi: 10.1016/j.jbusres.2015.07.001.

Ertel, W. (2017). Machine Learning and Data Mining, in, pp. 175–243. doi: 10.1007/978-3-319-58487-4_8.

Esuli, A. and Sebastiani, F. (2006). {SENTIWORDNET}: A Publicly Available Lexical Resource for Opinion Mining, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation ({LREC}{'}06)*. Genoa, Italy: European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf.

Fan, J. and Gu, J. (2003). Semiparametric estimation of Value at Risk, *The Econometrics Journal*, 6(2), pp. 261–290. doi: 10.1111/1368-423X.t01-1-00109.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, 35(2), p. 137–144. doi: 10.1016/j.ijinfomgt.2014.10.007.

Gautam, G. and Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis, in *2014 Seventh International Conference on Contemporary Computing (IC3)*. IEEE, p. 437–442. doi: 10.1109/IC3.2014.6897213.

Giambattista, Amati Giuseppe, A. *et al.* (2008). FUB, IASI-CNR and University of Tor Vergata at TREC 2008 Blog Track, *NIST Special Publication*.

Giudici, P., Mezzetti, M. and Muliere, P. (2003). Mixtures of products of Dirichlet process for variable selection in survival analysis, *Journal of Statistical Planning and Inference*, 111(1–2), p. 101–115. doi: 10.1016/S0378-3758(02)00291-4.

Godbole, N., Manjunath, S. and Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs Namrata, in *Conference: Proceedings of the International Conference on Weblogs and Social Media*.

Hemmatian, F. and Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis, *Artificial Intelligence Review*, 52(3), pp. 1495–1545. doi: 10.1007/s10462-017-9599-6.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. doi: 10.1145/1014052.1014073.

Hutto, C.J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for, *Eighth International AAAI Conference on Weblogs and Social Media*, pp.18. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

Kabir, N. and Carayannis, E. (2013). Big data, tacit knowledge and organizational competitiveness, *Journal of Intelligence Studies in Business*, 3(3), pp.54–62. doi: 10.37380/jisib.v3i3.76.

Karimi, A., Rossi, L. and Prati, A. (2021). AEDA: An Easier Data Augmentation Technique for Text Classification, *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pp .2748–2754. doi: 10.18653/v1/2021.findings-emnlp.234.

Li, X. *et al.* (2019). Exploiting bert for end-to-end aspect-based sentiment analysis_, *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings*, pp. 34–41. doi: 10.18653/v1/d19-5505.

Liu, B. *et al.* (1998). Integrating Classification and Association Rule Mining, *Knowledge Discovery and Data Mining*, pp.80–86. Available at: http://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf%5Cnhttp://www.aaai.org/Library/KDD/1998/kdd98-012.php%5Cnhttp://citeseer.ist.psu.edu/liu98integrating.html.

Liu, Y. *et al.* (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach, (1). Available at: http://arxiv.org/abs/1907.11692.

Ma, D. *et al.* (2017). Interactive Attention Networks for Aspect-Level Sentiment Classification, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. California: International Joint Conferences on Artificial Intelligence Organization, pp. 4068–4074. doi: 10.24963/ijcai.2017/568.

Manyika, J. *et al.* (2011). Big data: The next frontier for innovation, competition and productivity, *McKinsey Global Institute*, (June), pp.156. Available at: https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf.

Michael Steinbach, George Karypis and Vipin Kumar (2000). A Comparison of Document Clustering Techniques, *KDD workshop on text mining*, pp.1–2. Available at: https://www.bibsonomy.org/bibtex/210e5c1e3ff54d9dce505a231f8ae7b32/hotho.

Miller, G. A. (1995). WordNet: A Lexical Database for English, *Communications of the ACM*, 38(11), pp.39–41. doi: 10.1145/219717.219748.

Mohanty, A. K., Senapati, M. R. and Lenka, S. K. (2013). An improved data mining technique for classification and detection of breast cancer from mammograms, *Neural Computing and Applications*, 22(1), pp.303–310. doi: 10.1007/s00521-012-0834-4.

Phan, M. H. and Ogunbona, P. O. (2020). Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis, pp. 3211–3220. doi: 10.18653/v1/2020.acl-main.293.

Prathi, J. K., Raparthi, P. K. and Gopalachari, M. V. (2020). Real-Time Aspect-Based Sentiment Analysis on Consumer Reviews, *Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing*, pp. 801–810. doi: 10.1007/978-981-15-1097-7_67.

Provost, F. and Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making, *Big Data*, 1(1), pp. 51–59. doi: 10.1089/big.2013.1508.

Russom, P. (2021) *Big data analytics*, *A Closer Look at Big Data Analytics*.

Sangeetha, S. and Sreeja, A. K. (2015). No Science No Humans, No New Technologies No Changes "Big Data a Great Revolution", *International Journal of Computer Science and Information Technologies*, 6(4), pp. 3269–3274.

Sheela, L. J. (2016). A Review of Sentiment Analysis in Twitter Data Using Hadoop,

International Journal of Database Theory and Application, 9(1), pp. 77–86. doi: 10.14257/ijdta.2016.9.1.07.

Shim, J. P. *et al.* (2015). Big data and analytics: Issues, solutions, and ROI, *Communications of the Association for Information Systems*, 37(1), pp. 797–810. doi: 10.17705/1cais.03739.

Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P. (2017). Divergence of sentiment and stock market trading, *Journal of Banking & Finance*, 78, pp. 130–141. doi: 10.1016/j.jbankfin.2017.02.005.

Thelwall, M., Homsi, M. N. and Prabowo, R. (2009). Sentiment analysis: A combined approach Cite this paper Related papers SA2 vinodhini Manieniyan Sent iment Analysis and Sent iment Classificat ion using NLP IRJET Journal Mult i-Class Sent iment Analysis using a Hierarchical Logist ic Model Tree Approach.

Theodorou, T. I. *et al.* (2021). An AI-enabled stock prediction platform combining news and social sensing with financial statements, *Future Internet*, 13(6), pp. 1–22. doi: 10.3390/fi13060138.

Wei, Y.-C. *et al.* (2017). Informativeness of the market news sentiment in the Taiwan stock market, *The North American Journal of Economics and Finance*, 39, pp. 158–181. doi: 10.1016/j.najef.2016.10.004.

Weng, B., Ahmed, M. A. and Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources, *Expert Systems with Applications*, 79, pp. 153–163. doi: 10.1016/j.eswa.2017.02.041.

Wu, X. *et al.* (2019). Conditional BERT Contextual Augmentation, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11539 LNCS, pp.84–95. doi: 10.1007/978-3-030-22747-0_7.

Xu, C. *et al.* (2020). BERT-of-Theseus: Compressing BERT by Progressive Module Replacing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, p.7859–7869. doi: 10.18653/v1/2020.emnlp-main.633.

Xu, H. *et al.* (2020). DomBERT: Domain-oriented language model for aspect-based sentiment analysis, *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pp.1725–1731. doi: 10.18653/v1/2020.findings-emnlp.156.

Xu, H. *et al.* (2021). Understanding Pre-trained BERT for Aspect-based Sentiment Analysis, p.244–250. doi: 10.18653/v1/2020.coling-main.21.

Yi, J. *et al.* (2003). Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, in *Third IEEE International Conference on Data Mining*. IEEE Comput. Soc, pp. 427–434. doi: 10.1109/ICDM.2003.1250949.

Yıldırım, D. C., Toroslu, I. H. and Fiore, U. (2021). Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators, *Financial Innovation*. Springer Berlin Heidelberg, 7(1), pp. 1–36. doi: 10.1186/s40854-020-00220-2.

Zhang, X. *et al.* (2018). Improving stock market prediction via heterogeneous information fusion, *Knowledge-Based Systems*, 143, pp. 236–247. doi: 10.1016/j.knosys.2017.12.025.