mathematics

MDPI

Article

Using Market News Sentiment Analysis for Stock Market Prediction

Marian Pompiliu Cristescu [1], Raluca Andreea Nerisanu [1,*], Dumitru Alexandru Mara [1] and Simona-Vasilica Oprea [2]

[1] Faculty of Economic Sciences, Lucian Blaga University of Sibiu, 550324 Sibiu, Romania
[2] Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania
* Correspondence: raluca.nerisanu@ulbsibiu.ro

Abstract: (1) Background: Since the current crises that has inevitably impacted the financial market, market prediction has become more crucial than ever. The question of how risk managers can more accurately predict the evolution of their portfolio, while taking into consideration systemic risks brought on by a systemic crisis, is raised by the low rate of success of portfolio risk-management models. Sentiment analysis on natural language sentences can increase the accuracy of market prediction because financial markets are influenced by investor sentiments. Many investors also base their decisions on information taken from newspapers or on their instincts. (2) Methods: In this paper, we aim to highlight how sentiment analysis can improve the accuracy of regression models when predicting the evolution of the opening prices of some selected stocks. We aim to accomplish this by comparing the results and accuracy of two cases of market prediction using regression models with and without market news sentiment analysis. (3) Results: It is shown that the nonlinear autoregression model improves its goodness of fit when sentiment analysis is used as an exogenous factor. Furthermore, the results show that the polynomial autoregressions fit better than the linear ones. (4) Conclusions: Using the sentiment score for market modelling, significant improvements in the performance of linear autoregressions are showcased.

Keywords: sentiment analysis; market prediction; BeautifulSoup; ARX; VADER

MSC: 62H12; 62J05; 68U15

check for updates

Citation: Cristescu, M.P.; Nerisanu, R.A.; Mara, D.A.; Oprea, S.-V. Using Market News Sentiment Analysis for Stock Market Prediction. *Mathematics* **2022**, *10*, 4255. https://doi.org/10.3390/math10224255

Academic Editors: Manuel Alberto M. Ferreira and Junseok Kim

Received: 22 September 2022
Accepted: 11 November 2022
Published: 14 November 2022

1. Introduction

Risk managers can consider a large set of assets, greater than the size of their portfolios, with the use of financial big data. In particular, when it comes to capturing the systematic risk built into markets, this usefulness can result in a higher accuracy of risk prediction. Therefore, in order to maintain the financial markets' stability and thereby lower the likelihood that a systemic risk materializes, financial regulators can benefit from a high level of accuracy in risk prediction for market participants.

The financial data of public companies are typically published on a relatively rare basis, which causes a clear time lag. Meanwhile, financial data tend to have a more frequent rate of appearance in financial news.

The approaches for portfolio weighting and stock selection are notably insensitive as a result of the rising popularity of high-frequency trading. In actuality, the information that investors can view in real time is stock trading data, such as the opening, highest, lowest and closing price of stock, as well as various technical indicators, and so on. In addition, investors can now use their decision sentiments based on text data found within financial news on the web, which can then be incorporated into stock-investment-value analysis, thanks to the growing availability of web-based data.

As asset prices incorporate many characteristics into their value, classical and modern approaches can be classified into two types, depending on the data that were modelled to predict asset prices. Thus, the fundamental approach may include such data as stock information parameters and "balance sheet & profit and loss statement parameters" [1]. Meanwhile, in [2], the two are grouped by company analysis, industry analysis [2], macroeconomic indicators [2], political circumstances [3] and geographical and meteorological circumstances [3]. Meanwhile, technical analysis refers to the analysis of prices [3], sentiment, raw data, volume, cycle, volatility, flow of funds [2] or other technical indicators [1].

While most of the technical and fundamental data are provided in a structured manner in the classical approaches, the modern approaches may perform on unstructured data sources, mainly reached through web-based financial news, social media, blogs, web-based forums and so on [4]. With an increasing number of websites and internet users, it can be challenging to locate and organize relevant information. Web scraping is the process of extracting information from a website by "scraping" it. Theoretically, it is feasible to scrape other data sources, such as document papers. Nonetheless, the vast majority of scraping is often performed on webpages.

As in behavioral economics, prices are purely a perceived value [5]. It is reasonable to search for the impact of society's opinions on the asset prices. This technique is called opinion mining, and it consists of identifying sentiments (positive or negative) through words.

Many academics now use investor sentiment to forecast stock price movement and portfolio optimization [6–8]. Pertinent studies on the use of online messaging to forecast stock market movements were compiled by [9] in their article. Their empirical findings imply that network messaging has some potential for applications in financial forecasting.

In order to examine the investment worth of equities, numerous academics have also started fusing together data from other sources. A number of stock-price-prediction experiments using numerous data sources have been conducted, including those by [10–13], among others. As far as we are aware, some academics have also suggested using big data to study stock selection and portfolio optimization, but the viability of this suggestion has not been proven (i.e., [7]). Therefore, our goal is to demonstrate how data science techniques may be used to identify stocks that are suitable for investment in a securities market with a lot of assets.

The recent literature proposes many stock-market-prediction techniques, useful for both technical and fundamental analysis. Machine learning algorithms include classification techniques (support vector machines, k-nearest neighbors, logistic regressions, naïve Bayes, decision tree classification and random forest classification), regression techniques (polynomial regression, simple linear regression, decision tree regression, random forest regression and support vector regression), fuzzy logic algorithms, deep neural networks, genetic algorithms and artificial neural networks [14–18]. In [15], two pattern-recognition techniques were found, namely template matching and perceptually important points (PIPs).

More phases were proposed in the machine-learning market-prediction process, which can be summarized by three phases, as Figure 1 shows:

1. Dataset preparation: social media, reports, blogs, online news or market data [3,4];
2. Data pre-processing [3]:
    - Feature selection (in reference to natural language processing),
    - Dimensionality reduction,
    - Feature representation;
3. Modelling and evaluation.

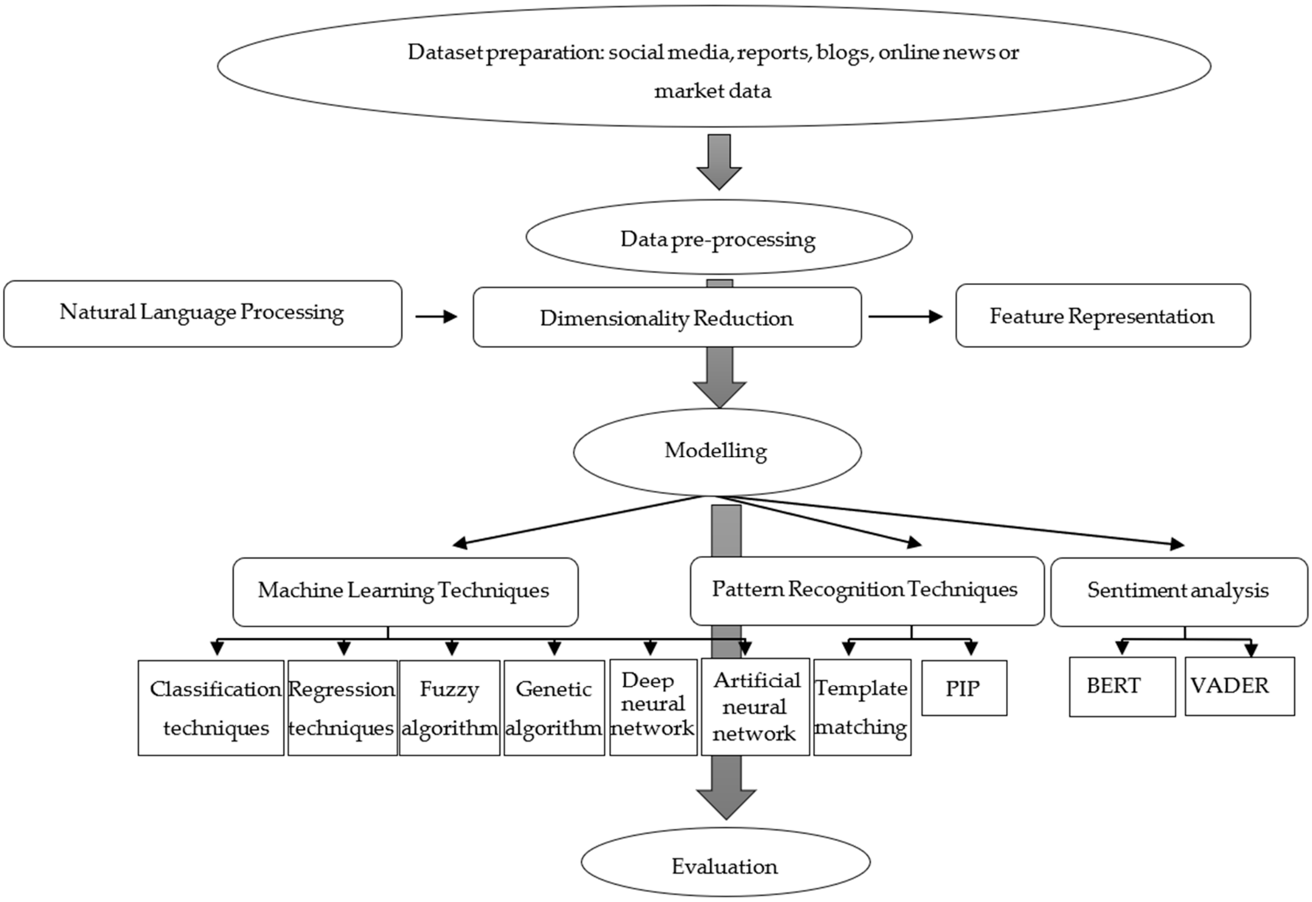
Dataset preparation: social media, reports, blogs, online news or market data
Data pre-processing
Natural Language Processing
Dimensionality Reduction
Feature Representation
Modelling
Machine Learning Techniques
Pattern Recognition Techniques
Sentiment analysis
Classification techniques
Regression techniques
Fuzzy algorithm
Genetic algorithm
Deep neural network
Artificial neural network
Template matching
PIP
BERT
VADER
Evaluation

Figure 1. Using machine learning in market prediction process.

In [3], feature selection techniques were classified as follows: bag of words; n-grams as continuous sequences of words; genetic algorithms; and colony optimization. Meanwhile, feature representation techniques may include information gain (IG), chi-square statistics (CHI), document frequency (DF), accuracy balanced (ACC2), term frequency–inverse document frequency (TF–IDF), binary/Boolean (0/1) or sentiment value.

The analysis of human opinions expressed in text is known as sentiment analysis and is one of the natural language processing (NLP) tasks [19]. Obtaining categories according to polarity (positive/negative/neutral expression), topic classification (determining the subjectivity or objectivity of an expression), and irony detection (determining whether a phrase is ironic) are the main objectives of sentiment analysis, based on different levels of granularity, such as documents, sentences or aspects. Thanks to the development of social networks and their usage in various industries, such as consumer goods and healthcare, financial sentiment analysis has a wide range of possible applications.

Sentiment analysis may usually include some feature representation techniques along with machine learning techniques (e.g., the bag-of-words model plus support vector machines [20]), but some models have been widely used, namely Pre-training of Deep Bidirectional Transformers (BERT) and the Valence Aware Dictionary for Sentiment Reasoning (VADER) [21–23].

The next section includes the methodological approach used in order to incorporate sentiment analysis in market predictions, while the research design is presented in Figure 2. To reach our purpose, we used an autoregression with an exogenous factor model (ARX), along with quadratic and cubic regressions. In the third section, the results are presented, starting with graphs that cover the sentiment score for each of the analyzed stock. Additionally, a scatter plot constructing the relationship between the sentiment score and the stock opening price is presented in Section 3.1. In Section 3.2, the regression models are presented and discussed, while the importance of integrating the sentiment score into these regressions is highlighted. In Section 4, the discussion of the results and a comparison with similar studies is introduced, while the last section focuses on the main conclusions.

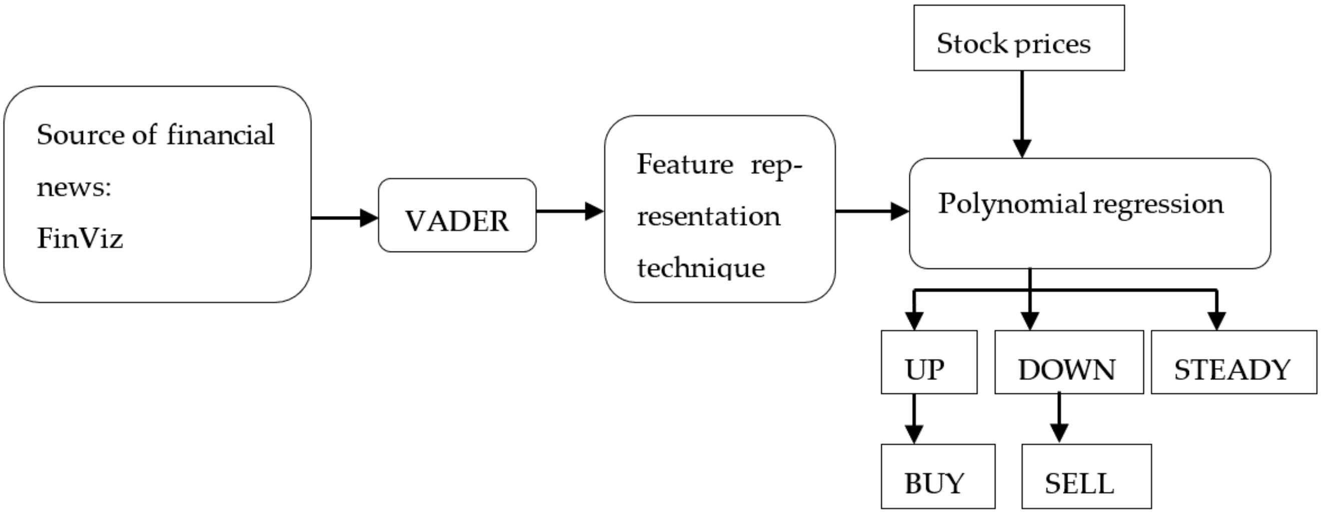

Figure 2. Research design.

2. Materials and Methods

In order to perform our analysis, we used the FinViz platform to gain financial news on various active stocks. We also made use of Python, which was the second-most popular programming language in 2020, behind C [24]. This programming language introduced in 1991 received subsequent reworks over the years, with the most significant upgrade occurring in 2008 with the release of Python 3.0. Python presents an extensive standard library with various useful tools, such as BeautifulSoup, a package that can be used for scraping and parsing website data [25].

After the selection of the data source, we ran a Python script that uses BeautifulSoup to scrape article headlines from FinViz, which is a platform for researching the stock market that is available through a web browser [26]. Afterwards, we used VADER to run the sentiment analysis, and we used Pandas (the Python data analysis library) to analyze and return the resultant sentiment analysis scores for the headlines of the financial articles.

BeautifulSoup is the most popular package for scraping and parsing website data. According to the makers of the library, it can interpret any input. BeautifulSoup does this by employing simple methods and Pythonic idioms to construct a navigable and searchable parse tree. The benefit of utilizing BeautifulSoup is that it translates parsed data to UTF-8, a widely used format on the internet [27]. The web scraper we used for data collection was created using tools available in the BeautifulSoup library.

In order to conduct sentiment analysis, it is necessary to apply a model. VADER (which stands for "Valence Aware Dictionary for Sentiment Reasoning") is a simple rule-based model for general sentiment analysis. This model is sensitive to both polarity and the strength of emotion, and it can be applied to unlabeled text data. VADER is included in the NLTK package, which represents a platform for the building of Python programs that enable working with human language data [28]. VADER shares the benefits of traditional sentiment lexicons, such as LIWC (linguistic inquiry and word count) and improves upon them. VADER differs from LIWC in that it generalizes more favorably to different domains and is more responsive to sentiment expressions in social media environments. Hutto and Gilbert were able to design and empirically validate a set of lexical characteristics that are particularly sensitive to sentiment in microblog-like circumstances. VADER performed as well as eleven other highly regarded sentiment analysis tools [22]. Therefore, VADER can be used for the sentiment analysis of headlines regarding financial news published in the online environment and shared on social media. However, it is important to mention that there were various dates in which the selected stocks were not covered by the major finance news publications that FinViz collects its data from; therefore, the sentiment score was 0.

After the sentiment score was obtained, in the modelling phase two linear autoregressions were performed; one without an exogenous factor and one with one (the exogenous factor being the sentiment score) as Equation (1) and Equation (2), respectively, show:

$$y = \beta_1 + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + \beta_5 y_{t-4} + \varepsilon \tag{1}$$

$$y = \beta_1 + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + \beta_5 y_{t-4} + \beta_6 x + \beta_7 x_{t-1} + \beta_8 x_{t-2} + \varepsilon \tag{2}$$

where β_1 is the intercept, β_2 is the slope, $y_{t-1/2/3/4}$ are the stock-opening-price change variables in the previous days, x and $x_{\mathrm{t}-1}$ are the sentiment score exogenous variables in the present and previous days and ε is the error term. y is the stock price change in time t.

Meanwhile linear, quadratic and cubic regressions were used in order to analyze the relation between the sentiment score and the stock-opening-price change. The relation for the linear, quadratic and cubic autoregressions are presented in Equation (3), Equation (4) and Equation (5), respectively:

$$y = \beta_1 + \beta_2 x + \varepsilon \tag{3}$$

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon \tag{4}$$

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \varepsilon \tag{5}$$

where y is the stock opening price change, β_1, β_2, β_3 and β_4 are the unknown parameters, x is the sentiment score variable and ε is the error term. Y is the stock price change in time t.

After running the linear, quadratic and cubic regressions, we ran a nonlinear autoregression with an exogenous factor (NARX) as can be seen in Equation (6):

$$y = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_4 y_{t-4} + \beta_5 x + \beta_6 x^2 + \varepsilon \tag{6}$$

where y is the opening price change in time t, x is the sentiment score and $y_{\text{t-1/2/3/4}}$ are the opening price changes in time $t-1$, $t-2$, $t-3$ and $t-4$.

In order to perform our regression, we used ordinary least squares in SPSS.

We also used aggregated data in order to perform linear autoregressions with and without an exogenous factor, in which the sentiment score was aggregated with the market capitalization weight. The results show that the models with of equal weight had better determinant coefficients for the training data; therefore, we maintained an equal weight for all of the regressions included in the present study.

The time period analyzed covers data from August–September 2022, consisting of a total of 37 days. All data were aggregated using equal weight.

Most of the companies engaged in the study were selected from the S&P 500 due to their popularity, as well as the fact that any investor who is interested in the stock market and invests in this index is exposed to them. One company, Unity Technologies, was selected due to its recent IPO (initial public offering) on 17.09.2020 [29]. Additionally, we included a company that it is not based in United States, SONY, to observe the difference in sentiment scores of the news of companies within and beyond the United States. The market data from the companies were extracted from The New York Stock Exchange.

3. Results

3.1. Sentiment Analysis

After scraping the text from FinViz for each selected company and analyzing the sentiment of the articles from August and September 2022, we calculated an average sentiment score in the news of 0.06 and an average volatility of 0.18. For each stock, the sentiment scores are presented in Figure 3, from which we can observe the following aspects.

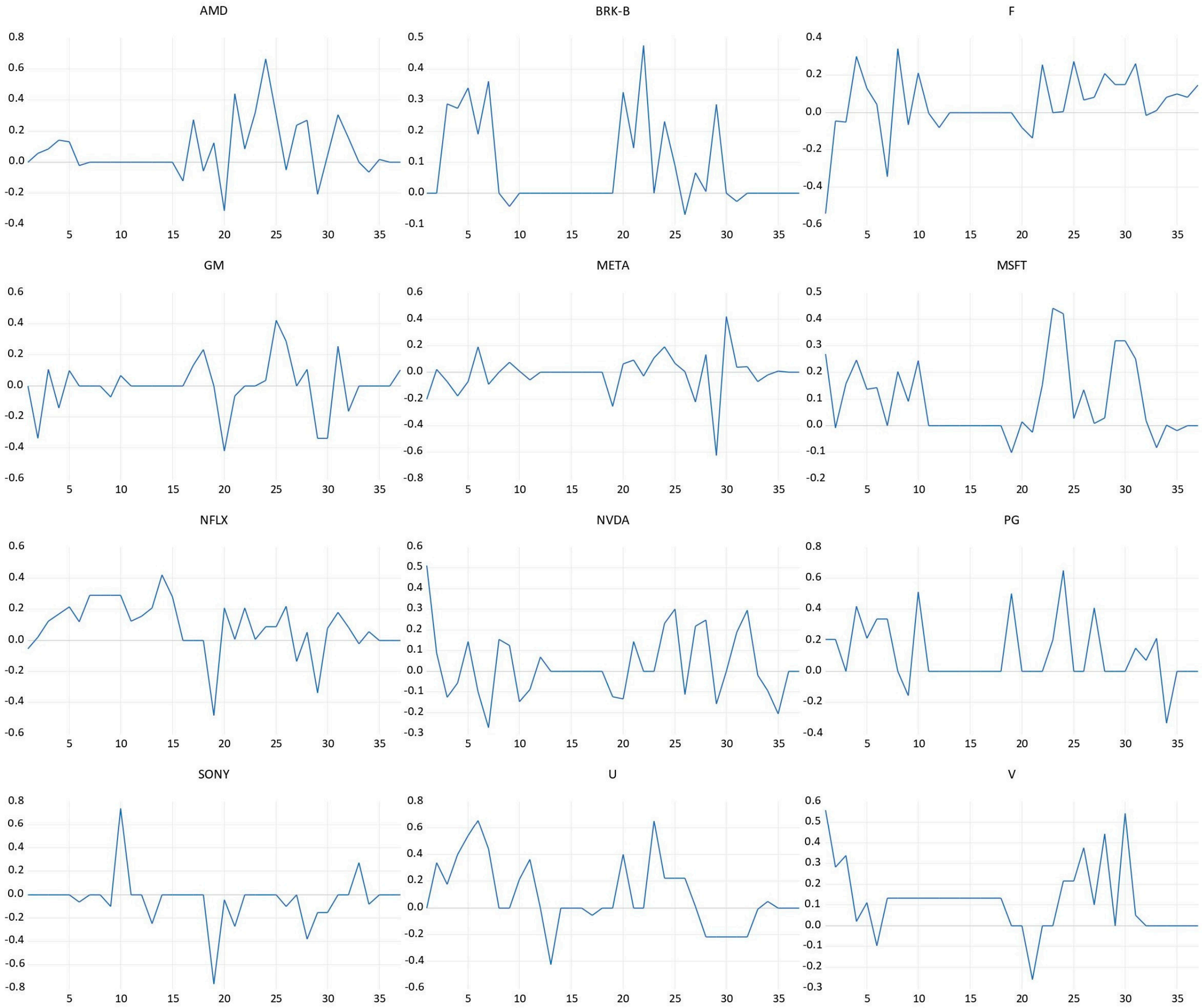

Figure 3. Sentiment analysis charts for each analyzed stock between 9 August 2022–30 September 2022.

The average sentiment score for AMD (Advanced Micro Devices) stock was a positive one with a value of 0.08, a value above total average, and a volatility of 0.18.

On average, the sentiment score in the news related to BRK-B (Berkshire Hathaway Inc. Class B) stock was 0.11, indicating a positive sentiment in the news related to this stock. The volatility score was 0.202, a relatively high one, which was above the average volatility of 0.18. Additionally, for F (Ford Motor Company) stock, the average sentiment score was 0.04, indicating a positive sentiment in the news related to this stock, and it had a below-average volatility of 0.17. As for MSFT (Microsoft Corp.), a positive sentiment was seen in relation to stock, with an average sentiment score of 0.09 and the lowest volatility of 0.14 during the analyzed period. The average sentiment scores for GM (General Motors) and META (Meta, Inc., formerly Facebook, Inc.) stocks were −0.001 and −0.001, indicating negative sentiments in the news related to these stocks, both having a volatility of 0.17 and 0.16. Meanwhile, while for NFLX (Netflix) stock, the average sentiment score was 0.09, indicating a positive sentiment in the news related to this stock, and volatility was 0.14 during the analyzed period, below the average volatility of 0.18. Similar to MSFT, news related to PG (Procter & Gamble) had a positive sentiment score of 0.11 and a volatility score of 0.20, which was above average.

For NVDA (Nvidia), the average sentiment score was 0.03, indicating a positive sentiment in the news related to this stock, and it had a volatility of 0.16, which was relatively low. V (Visa) stock-related news had the highest positive sentiment score of 0.12 with a volatility of 0.17. Meanwhile, the average sentiment score for U (Unity Technologies) was a positive one with a value of 0.09 and it had the highest volatility, one of 0.25. The news related to SONY had the most negative sentiment score with a value of −0.04, while its volatility stood above average with a value of 0.20.

Regarding the opening price volatility, Figure 4 presents the period volatility for each analyzed stock.

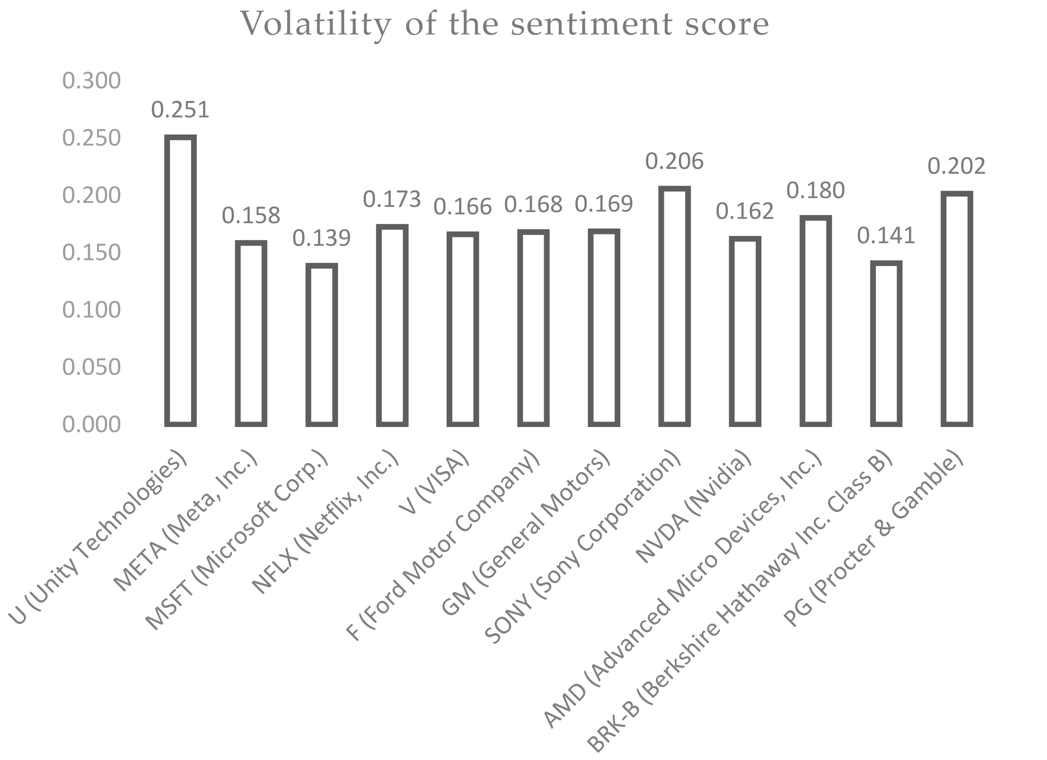

Figure 4. Stock price volatility in the analyzed period for the opening price.

While the highest standard deviation (volatility) was 0.25 for the U (Unity) stock, the lowest volatility stood for MSFT (Microsoft Corp.) stock at 0.14. In the short term, for AMD stock, the opening price of the market increased on 19.08.2022 from that of 18.08.2022 as it presented the highest sentiment score for Advanced Micro Devices (AMD) of 0.66 on 19.08.2022 and the highest average sentiment score of 0.12 during the aforementioned period. This may suggest that the news sentiment may have affected this change.

3.2. Regression Models

In order to predict and forecast the opening price change, we proposed a linear autoregression, without and with an exogenous factor. The linear autoregression without the exogenous factor had an R of 0.189 as Table 1 shows, while adding the sentiment score as an exogenous factor raised the R coefficient to 0.192, as Table 2 shows.

Table 1. Linear autoregression.

Model 1 Summary [a,b]		
R	R Square	Adjusted R Square
0.189	0.036	0.024

a. Predictors: (constant), opening price change $t-1$, opening price change $t-2$, opening price change $t-3$, opening price change $t-4$. b. Dependent variable: opening price change t.

Table 2. Linear autoregression with exogenous factor.

Model 2 Summary [a,b]		
R	R Square	Adjusted R Square
0.192	0.037	0.014

a. Predictors: (constant), sentiment score *t*, sentiment score $t-1$, sentiment score $t-2$, opening price change $t-1$, opening price change $t-2$, opening price change $t-3$, opening price change $t-4$. b. Dependent Variable: opening price change *t*.

The model of linear autoregression with an exogenous factor is presented in Table 3. As one unit grows in opening price change, time $t-2$ predicts a 0.124 change in the next day's opening price change, while a one-unit growth in sentiment score predicts 0.063 units growth in the opening price change of the next day. From the comparison of the unstandardized coefficients, when applying the model to the testing data, we can observe that the coefficients maintained the same sign and approximatively the same value, except for the Beta coefficient for the opening price change of the previous day (change $t-1$) and for the sentiment score of the previous day (sentiment $t-1$).

Table 3. Linear autoregression with exogenous factor coefficients.

		Coefficients for Training Data [a]					Coefficients for Test Data [a]	
Model		**Unstandardized Coefficients**		**Standardized Coefficients**	***t***	**Sig.**	**Unstandardized Coefficients**	**Sig.**
		B	**Std. Error**	**Beta**				
2	β_1 (Constant)	−0.347	0.121		−2.881	0.004	0.079	0.857
	β_2 (Change $t-1$)	0.067	0.057	0.069	1.188	0.236	−0.151	0.130
	β_3 (Change $t-2$)	0.124	0.060	0.122	2.086	0.038	0.092	0.270
	β_4 (Change $t-3$)	0.073	0.061	0.070	1.196	0.233	0.082	0.304
	β_5 (Change $t-4$)	0.028	0.061	0.027	0.461	0.645	0.190	0.015
	β_6 (Sentiment *t*)	0.063	0.534	0.007	0.117	0.907	0.655	0.726
	β_7 (Sentiment $t-1$)	0.000	0.567	0.000	0.001	0.999	0.764	0.711
	β_8 (Sentiment $t-2$)	−0.547	0.586	−0.055	−0.932	0.352	−3.184	0.245

a. Dependent variable: opening price change *t*.

Additionally, ANOVA analysis was applied to the two autoregressions, the results of which are shown in Table 4. After running the ANOVA, the F test was applied to compare the goodness of fit of the two models.

Table 4. Goodness of fit analysis—ANOVA.

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
1	Regression	37.483	4	9.371	2.949	0.020
	Residual	1013.585	319	3.177		
	Total	1051.068	323			
2	Regression	35.168	7	5.024	1.597	0.136
	Residual	918.892	292	3.147		
	Total	954.060	299			

1: Predictors: opening price change $t-1$, opening price change $t-2$, opening price change $t-3$, opening price change $t-4$. 2: Predictors: sentiment score *t*, sentiment score $t-1$, sentiment score $t-2$, opening price change $t-1$, opening price change $t-2$, opening price change $t-3$, opening price change $t-4$.

The F statistic was calculated in order to analyze the goodness of fit of the two models. The F statistic computed for the models with a different number of parameters was 1.1144, while the computed *p*-value was estimated to be closer to 0 (0.1729); thus, the first model was more fitted than the second one, although the sum of squares is larger in the first model.

Other types of regressions among the stock opening price change and the sentiment score are presented in Table 5.

Table 5. Different regressions among opening price and sentiment score.

Model Summary and Parameter Estimates [a]									
	Dependent variable: opening price change								
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	β_1	β_2	β_3	β_4
Linear	0.001	0.478	1	346	0.49	0.004	0.004		
Quadratic	0.004	0.718	2	345	0.488	0.004	0.002	0.013	
Cubic	0.005	0.553	3	344	0.647	0.004	−0.001	0.011	0.014

a. The independent variable is the sentiment score.

As Table 5 shows, the R squared value in the computed regressions, which used the opening price as the dependent variable and the sentiment score as the independent variable, shows that the polynomial regressions among the two variables were more suited than the linear one. The coefficients of the quadratic regression are presented in Equation (7) and those of the cubic regression are shown in Equation (8):

$$y = 0.004 + 0.002 \cdot x + 0.013 \cdot x^2 + \varepsilon \tag{7}$$

$$y = 0.004 - 0.001 \cdot x + 0.011 \cdot x^2 + 0.14 \cdot x^3 + \varepsilon \tag{8}$$

As the quadratic and cubic models were the best fitted ones, a nonlinear autoregressive model with exogenous factor is suitable for the available data, combining the two models as shown in Table 6.

Table 6. Estimators for the NARX model.

Parameter Estimates				
Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
β_1 (Change $t-1$)	0.108	0.071	−0.032	0.247
β_2 (Change $t-2$)	0.157	0.071	0.017	0.297
β_3 (Change $t-3$)	0.139	0.009	0.122	0.156
β_4 (Change $t-4$)	0.050	0.069	−0.085	0.186
β_5 (Sentiment t)	0.008	0.019	−0.029	0.045
β_6 (Sentiment $t^{2)}$	0.016	0.070	−0.122	0.154

The R squared value of the model was 0.005, as shown in Table 7.

Table 7. ANOVA analysis for the NARX.

ANOVA [a]			
Source	Sum of Squares	df	Mean Squares
Regression	0.010	9	0.001
Residual	0.087	219	0.000
Uncorrected total	0.097	228	
Corrected total	0.091	227	
Dependent variable: Op_pricet			

a. R squared = 1; (Residual Sum of Squares)/(Corrected Sum of Squares) = 0.005.

4. Discussion

In the present paper, we used the FinViz platform to obtain financial news headlines on selected stocks from popular financial news websites, followed by the application of the VADER model in order to foster the general sentiment toward events that can occur regarding the analyzed stocks. The model involves a Python script that uses BeautifulSoup to scrape article headlines from FinViz, and Pandas to analyze and return the resultant sentiment analysis scores for the headlines of the financial articles.

Future stock trend analysis is a challenging endeavor due to the multiplicity of variables involved. We hypothesized that news items and stock prices are correlated, and that the news may correspond with the swings of stock prices.

Sentiment analysis was conducted daily for the analyzed period by collecting the headlines of the news from FinViz and applying the VADER model in Python to obtain the sentiment scores. As we can see from the results, the sentiment scores varied significantly from one day to another. The average sentiment of the market news between 06.08.2022 and 30.09.2022 was 0.06, which indicates a positive sentiment evoked by the news. The lowest sentiment score was −0.765 for Sony Corporation on 24.08.2022 and the highest sentiment score was 0.743 on 22.09.2022.

For the SONY stock, the opening price of the market decreased on 24.08.2022 from that of 23.08.2022. Additionally, for the AMD stock, the opening price of the market increased on 19.08.2022 from that of 18.08.2022. This suggest that news sentiment may have affected this change. After calculating the volatility of the sentiments for these scores, we identified U (Unity Technologies) stock as having the highest volatility in its sentiment score and MSFT (Microsoft Corp.) stock as having the lowest volatility. This can contribute to the idea that MSFT (Microsoft Corp.) stock is a less volatile investment, given that its news headlines are relatively consistent in their sentiment and the opinions of major finance publications about this company are not very divided.

The regressions that we used were cubic, quadratic and linear regressions, as the cubic regressions were found to have a higher accuracy than the linear one [14]. Our results show that the polynomial regressions were more fitted with the model than the linear one, as the R squared value for the cubic regression was 0.005, while for the linear one, this value was 0.001. Additionally, in [14], the polynomial regression accuracy was only surpassed by the decision tree type of regression. Furthermore, from the decision tree regressions, random forest regression had the highest accuracy [14,30].

Similar results, such as those presented in [31], propose an algorithm that combines the price indices of the analyzed stock with the daily sentiment of each stock and recommends "additional signals" based on the analyzed sentiment. By using two individual "long short-term memories" that were merged through smart decision logic in [32], a profit accuracy for the one-to-five day FOREX profit of 63.91–73.09% was found. In [33], a polarized investor sentiment was found to be more determinant into market speculative bubbles than in the general volume of news and Google queries.

In another study, based on financial news published in China pertaining to companies listed on the Taiwan Stock Exchange, a market "Aggregate News Sentiment Index" (ANSI) was created and used to examine the correlation between the ANSI and market reactions [34]. In [35], a good association was found between sentiment disagreement and stock price volatility using data collected from Facebook regarding status updates to evaluate the gap between positive and negative feelings that occur daily in 20 countries. In order to assess the price and volume movement of stock on the following trading day, some studies were constructed upon four separate datasets (such as news from Google News, Wikipedia's trade information regarding business pages, typical technical indicators and historical stock trading data) [36]. The findings demonstrate that expanding the number of data sources can enhance forecast precision. In order to study the combined effects of many information sources on stock price movements, coupling matrices and tensor decomposition were employed in [37]. They also exploited the commonality between stocks to predict the price movements of numerous connected stocks at once.

Future studies may focus on expanding the types of data sources in order to raise the precision of market predictions, as [36] shows, or use combined techniques in order to reach better predictions. Additionally, nonparametric models may be used to enhance more accurate behavior of the stock market (e.g., random forest or decision tree regression) [33].

5. Conclusions

In the present paper, the sentiment factor was used to raise the goodness of fit for the prediction of the stock prices using regression models. We used the VADER model to generate the sentiment score that occurred each day, based on the financial news headlines that FinViz provided for selected stocks.

Next, we performed three types of regression models (linear, quadratic and cubic autoregressions), and found that the polynomial autoregressions had a higher R square indicator than the linear one, supporting the findings from [14]. Additionally, in order to improve the goodness of fit of the autoregression we used the sentiment score as an exogenous factor in nonlinear autoregression.

Our results are consistent with [14,30,35], while improving the evidence that including sentiment analysis as an exogenous factor in regression-type models can raise the goodness of fit of the models.

Overall, the results show that using the sentiment score as an exogenous factor in the linear autoregression raised the R coefficient from 0.189 to 0.192; thus, we can conclude that integrating the sentiment factor in market regression analysis generates a better regression regarding the goodness of fit.

Author Contributions: Conceptualization, M.P.C., R.A.N., D.A.M. and S.-V.O.; methodology, M.P.C., R.A.N., D.A.M. and S.-V.O. ; software, M.P.C., R.A.N., D.A.M. and S.-V.O. ; validation, M.P.C., R.A.N., D.A.M. and S.-V.O. ; formal analysis, M.P.C., R.A.N., D.A.M. and S.-V.O. ; investigation, M.P.C., R.A.N., D.A.M. and S.-V.O. ; resources, M.P.C., R.A.N., D.A.M. and S.-V.O. ; data curation, M.P.C., R.A.N., D.A.M. and S.-V.O. ; writing—original draft preparation, M.P.C., R.A.N., D.A.M. and S.-V.O. ; writing—review and editing, M.P.C., R.A.N., D.A.M. and S.-V.O. ; visualization, M.P.C., R.A.N., D.A.M. and S.-V.O. ; supervision, M.P.C., R.A.N., D.A.M. and S.-V.O. ; project administration, M.P.C., R.A.N., D.A.M. and S.-V.O. ; funding acquisition, M.P.C., R.A.N., D.A.M. and S.-V.O. All authors have read and agreed to the published version of the manuscript.

Funding: This project was financed by Lucian Blaga University of Sibiu through the research grant LBUS-IRG-2022-08. The APC was funded by Lucian Blaga University of Sibiu.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine Learning Techniques and Data for Stock Market Forecasting: A Literature Review. *Expert Syst. Appl.* **2022**, *197*, 116659. [CrossRef]
2. Hu, Y.; Liu, K.; Zhang, X.; Su, L.; Ngai, E.W.T.; Liu, M. Application of Evolutionary Computation for Rule Discovery in Stock Algorithmic Trading: A Literature Review. *Appl. Soft Comput.* **2015**, *36*, 534–551. [CrossRef]
3. Bustos, O.; Pomares-Quimbaya, A. Stock Market Movement Forecast: A Systematic Review. *Expert Syst. Appl.* **2020**, *156*, 113464. [CrossRef]
4. Nti, I.K.; Adekoya, A.F.; Weyori, B.A. *A Systematic Review of Fundamental and Technical Analysis of Stock Market Predictions*; Springer: Dordrecht, The Netherlands, 2020; Volume 53, ISBN 0123456789.
5. Nagy, J. *Behavioral Economics and the Effects of Psychology on the Stock Market*; State University of New York College at Buffalo—Buffalo State College: Buffalo, NY, USA, 2017.
6. Hochreiter, R. Computing Trading Strategies Based on Financial Sentiment Data Using Evolutionary Optimization. In *Advances in Intelligent Systems and Computing*; Springer International Publishing: New York, NY, USA, 2015; pp. 181–191. ISBN 9783319198248.
7. Jothimani, D.; Shankar, R.; Yadav, S.S. A Big Data Analytical Framework for Portfolio Optimization. *Cell* **2014**, *3*, 1–15.
8. Creamer, G.G. Can a Corporate Network and News Sentiment Improve Portfolio Optimization Using the Black–Litterman Model? *Quant. Financ.* **2015**, *15*, 1405–1416. [CrossRef]
9. Nardo, M.; Petracco-Giudici, M.; Naltsidis, M. Walking down wall street with a tablet: A survey of stock market predictions using the web. *J. Econ. Surv.* **2016**, *30*, 356–369. [CrossRef]
10. Shynkevich, Y.; McGinnity, T.M.; Coleman, S.A.; Belatreche, A. Forecasting Movements of Health-Care Stock Prices Based on Different Categories of News Articles Using Multiple Kernel Learning. *Decis. Support Syst.* **2016**, *85*, 74–83. [CrossRef]
11. Feuerriegel, S.; Gordon, J. Long-Term Stock Index Forecasting Based on Text Mining of Regulatory Disclosures. *Decis. Support Syst.* **2018**, *112*, 88–97. [CrossRef]
12. Atkins, A.; Niranjan, M.; Gerding, E. Financial News Predicts Stock Market Volatility Better than Close Price. *J. Financ. Data Sci.* **2018**, *4*, 120–137. [CrossRef]

13. Zhou, Z.; Gao, M.; Liu, Q.; Xiao, H. Forecasting Stock Price Movements with Multiple Data Sources: Evidence from Stock Market in China. *Phys. A Stat. Mech. Its Appl.* **2020**, *542*, 123389. [CrossRef]
14. Ravikumar, S.; Saraf, P. Prediction of Stock Prices Using Machine Learning (Regression, Classification) Algorithms. In Proceedings of the 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 5–7 June 2020; pp. 1–5. [CrossRef]
15. Shah, D.; Isah, H.; Zulkernine, F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *Int. J. Financ. Stud.* **2019**, *7*, 26. [CrossRef]
16. Rouf, N.; Malik, M.B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H.C. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. *Electronics* **2021**, *10*, 2717. [CrossRef]
17. Usmani, S.; Shamsi, J.A. News Sensitive Stock Market Prediction: Literature Review and Suggestions. *PeerJ Comput. Sci.* **2021**, *7*, e490. [CrossRef] [PubMed]
18. Ferreira, F.G.D.C.; Gandomi, A.H.; Cardoso, R.T.N. Artificial Intelligence Applied to Stock Market Trading: A Review. *IEEE Access* **2021**, *9*, 30898–30917. [CrossRef]
19. Nasukawa, T.; Yi, J. Sentiment Analysis. In Proceedings of the International Conference on Knowledge Capture—K-CAP '03; ACM Press: New York, NY, USA, 2003; p. 70.
20. Schumaker, R.P.; Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News. *ACM Trans. Inf. Syst.* **2009**, *27*, 1–19. [CrossRef]
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MA, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
22. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for sentiment analysis of social media text. In Proceedings of the international AAAI Conference on Web and Social Media, Catalonia, Spain, 17–21 July 2014; p. 18.
23. Leow, E.K.W.; Nguyen, B.P.; Chua, M.C.H. Robo-Advisor Using Genetic Algorithm and BERT Sentiments from Tweets for Hybrid Portfolio Optimisation. *Expert Syst. Appl.* **2021**, *179*, 115060. [CrossRef]
24. Tung, L. Programming Language Pythons' Popularity Ahead of Java for the First Time but Still Trailing c. Available online: https://www.zdnet.com/article/programming-language-pythons-popularity-ahead-of-java-for-first-time-but-still-trailing-c/ (accessed on 12 July 2022).
25. Breuss, M. Beautiful Soup: Build a Web Scraper with Python. *Preuzeto* **2021**, *30*, 2021.
26. FINVIZ. Available online: https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/finviz/ (accessed on 14 January 2022).
27. Beautiful Soup Documentation. Available online: https://beautiful-soup-4.readthedocs.io/en/latest/ (accessed on 12 July 2022).
28. Natural Language Toolkit. Available online: https://www.nltk.org/ (accessed on 12 July 2022).
29. Unity Software Inc. Financials. Available online: https://www.crunchbase.com/organization/unity-technologies/company_financials (accessed on 12 July 2022).
30. Sadorsky, P. A Random Forests Approach to Predicting Clean Energy Stock Prices. *J. Risk Financ. Manag.* **2021**, *14*, 48. [CrossRef]
31. Theodorou, T.I.; Zamichos, A.; Skoumperdis, M.; Kougioumtzidou, A.; Tsolaki, K.; Papadopoulos, D.; Patsios, T.; Papanikolaou, G.; Konstantinidis, A.; Drosou, A.; et al. An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements. *Futur. Internet* **2021**, *13*, 138. [CrossRef]
32. Yıldırım, D.C.; Toroslu, I.H.; Fiore, U. Forecasting Directional Movement of Forex Data Using LSTM with Technical and Macroeconomic Indicators. *Financ. Innov.* **2021**, *7*, 1–36. [CrossRef]
33. Giudici, P.; Mezzetti, M.; Muliere, P. Mixtures of Products of Dirichlet Process for Variable Selection in Survival Analysis. *J. Stat. Plan. Inference* **2003**, *111*, 101–115. [CrossRef]
34. Wei, Y.-C.; Lu, Y.-C.; Chen, J.-N.; Hsu, Y.-J. Informativeness of the Market News Sentiment in the Taiwan Stock Market. *North Am. J. Econ. Financ.* **2017**, *39*, 158–181. [CrossRef]
35. Siganos, A.; Vagenas-Nanos, E.; Verwijmeren, P. Divergence of Sentiment and Stock Market Trading. *J. Bank. Financ.* **2017**, *78*, 130–141. [CrossRef]
36. Weng, B.; Ahmed, M.A.; Megahed, F.M. Stock Market One-Day Ahead Movement Prediction Using Disparate Data Sources. *Expert Syst. Appl.* **2017**, *79*, 153–163. [CrossRef]
37. Zhang, X.; Zhang, Y.; Wang, S.; Yao, Y.; Fang, B.; Yu, P.S. Improving Stock Market Prediction via Heterogeneous Information Fusion. *Knowl.-Based Syst.* **2018**, *143*, 236–247. [CrossRef]