DATA ACQUISITION AND PROCESSING SYSTEMS ELEC0136 20/21 REPORT

SN: 20172994

ABSTRACT

In this report, the prediction of the Microsoft stock value is investigated starting from the data acquisition phase up to the inference process.[1] The task is addressed pursuing two distinct strategies according to the data exploited by the model selected to perform the forecasting. In the first approach, the information retrieved is obtained exclusively from the company's stock data. In the second strategy, the model leverages also the overall public sentiment towards the company, extracted from the Twitter platform, along with the records related to the ongoing pandemic. The experimental results obtained show that the already satisfactory performance achieved with the sole stock data analysis can be further enhanced through additional information that could influence the stock prices.

Index Terms— stock market prediction, twitter, sentiment analysis, natural language processing, mood analysis, pandemic, arima, facebook prophet, time series forecasting

1 Introduction

Stock market prediction represents an area of ever-growing interest that has lately attracted innumerable researchers from the academia as well as from the business world. Regardless of the prediction is made by individuals, by investment companies or by listed companies, the capability of understanding stock market movements in advance enables to operate promptly in order to increase a profit or reduce a loss. Early works in this field, such as [1, 2], firmly supported the Efficient Market Hypothesis (EMH) and the random walk theory. As described by Hellstrom et al. in [3], the strong form of EMH states that stock market price always reflects the entire information available, whether it is public or private. Whenever new information is released it is almost instantly assimilated by the system through a specific adjustment in the market price. Furthermore, according to the random walk theory, the stocks value is characterised by identically distributed and mutually independent fluctuations that lead to an entirely random and unpredictable movement. Both the theories argue the impossibility of predicting future price changes. Nonetheless, subsequent published studies, such as [4, 5], differed from the previous established viewpoints demonstrating effectively that stock market could to a certain extent be predicted. Albeit the first forecast attempts relied solely on historical stock data, a larger and more diversified amount of information can result into a more precise foresight capturing more details about the stock time-series. For instance, Schumaker et al. [6] showed the influence that breaking financial news have on stock variations predicting price movements few minutes after their release. More in general, it may stand to reason that public sentiment can direct stock variations just like news. The increasing presence of social media in the human daily routine makes these platforms the perfect place to comprehend the overall public opinion. Bollen et al. in [7], as well as other authors [8, 9], highlighted the existence of a meaningful correlation between the public mood extracted from Twitter posts, also known as tweets, and the Dow Jones Industrial Average (DJIA) values. Undoubtedly, according to the listed companies investigated, other information, about for instance the climate or the general health and well-being standards, may be useful to enhance the forecast reliability. In this work, Microsoft stock closing prices on each day of May 2020 are predicted starting from data spanning since the beginning of the fiscal year 2017 until the end of April 2020. In particular, the prediction is addressed pursuing two different approaches. Whereas in the first, the forecast is performed exploiting exclusively the company stock historical data, in the second, the model selected performs the task leveraging also additional information. In particular, the public sentiment towards the company is extracted from the Twitter platform and rendered to the model along with data regarding the ongoing pandemic caused by the proliferation of the coronavirus SARS-CoV-2. The rest of this report is organised as follows: an overview of the datasets used is reported in Section 2; a brief description on how data are acquired and collected is presented in Section 3; the storage strategies are explained in Section 4; an illustration of the preprocessing phase is proposed in Section 5; the data exploration process is outlined in Section 6; the results of the models adopted along with the experimental analysis are described in Section 7; and the main conclusions are summarised in Section 8.

2 Data description

As previously described, the stock data used to perform the predictions range from April 2017 to the end of April 2020. Although both the settings explored rely on the historical company's stock time-series, in the second case they are employed together with supplementary information extracted

[1] The code is entirely provided within the GitHub project repository reachable through the hyper-link: DAPS_assignment20_21.

from Twitter posts and from official daily reports of the current pandemic. If various studies cited in the introduction (Sec. 1) showcased that social platforms are valuable forecasting sources, the same is true for the reports on the virus outbreak that in 2020 has profoundly driven the global economy becoming a key element to be considered in the entire prediction process. In particular, the financial collapse experienced in March 2020 corresponds to one of the most significant crashes in the stock market history [10]. Hereafter, for each dataset, a description is provided with specific focus on the data measured from April 2017 up to April 2020.

2.1 Stock data & Indexes

The stock dataset consists of the union, realised during the preprocessing phase, of two different datasets. The first regards exclusively the company chosen and comprises 775 samples associated to each day inside the time interval selected whereby the National Association of Securities Dealers Automated Quotation (NASDAQ) market was operative. For each record, daily open, high, low, close, volume and adjusted close values are reported along with the historical split and dividend events. In particular, whereas the volume represents the amount of shares traded within a day, split and dividend events describe a stock price movement that is not directly captured by the close attribute and that is caused respectively by the division of individual stocks into smaller units and the distribution of dividends to shareholders. Finally, the adjusted closing price provides a better overview of the overall stock price since it reflects also the movement of the stock value caused by splittings and dividend payments. On the other side, the second dataset is always tied to the Microsoft company but it also concerns the other most important listed companies. It is made up of 775 observations defined by the closing values of the Standard and Poor (S&P) and DJIA indexes. Specifically, they report the stock performance of the major 500, 100 and 30 listed companies.

2.2 Tweets data

Given that a tweet has size constraints of solely 140 characters, it is consequently necessary to gather a massive amount of individual posts to properly identify the overall public mood towards the company. The dataset collected is made up of $234,523$ tweets defined by 36 attributes that report useful information such as the date of the post, the language used, the username of the submitter or the content published along with the inserted cashtags, hashtags and mentions. Specifically, cashtags are tags related exclusively to listed companies. The number of retweets, likes and replies for each tweet is provided as well.

2.3 News data

For the several reasons described in 3.3, the news dataset is collected starting from article headlines and arguments posted on Twitter by authoritative Anglo-Saxon financial sources. It consists of $12,111$ tweets characterised by the attributes of the dataset featured in 2.2.

2.4 Pandemic data

The dataset related to the contemporary health emergency relies on 100 observations made from January 22^{nd}, 2020 for each of the 199 countries taken into account for a total of $19,900$ records. Every sample is described by 35 attributes that allow to understand for instance where the measurement took place (country), the number of people that from the beginning of the pandemic contracted the virus (confirmed), were recovered (recovered) or passed away (deaths).

3 Data acquisition

The philosophy pursued during this phase consists in trying as far as possible to access databases through the official application program interfaces (APIs) available. Albeit they permit to accommodate the database owner policies, often APIs are content and/or rate limited. In such cases, data are collected using web scraping, i.e. to extract data parsing HTML webpages code.

3.1 Stock data & Indexes

Google Finance and Yahoo Finance were among the first web services to provide APIs to retrieve stock market information. Nevertheless, their interfaces did not receive support for few years becoming deprecated and today providing scarce documentation and reliability. A better solution adopted in this work is embodied by the Alpha Vantage free API [11] that gives access to real time and historical financial data going back up to 20 years. It also offers a practical python interface [12] through which queries can be easily sent.

However, although it enables the acquisition of data related to the sole company, it does not provide indexes information. Consequently, data associated to the second dataset are gathered scraping Yahoo Finance website via the pandas-datareader python package [13].

3.2 Tweets data

The free version of Twitter official API only enables to retrieve a restricted number of tweets from the last seven days. To overcome these limitations, the dataset is originated by means of a web scraping tool called Twint [14]. Hence, the tweets posted within the period observed are filtered and collected so that: they are written in English, they contain the cashtag keyword $MSFT and the related replies are not considered.

3.3 News data

Despite there are several APIs to access financial news databases, all of them have constraints similar to the Twitter API presented before. To avoid scraping several websites, the news dataset is created collecting article headlines and posts published on the Twitter platform by a series of hand-picked newspapers, economists and financial bloggers. In particular, the sources are mainly selected according to the number of followers and the language used, i.e. English. Finally, solely the tweets containing the keyword 'Microsoft' are gathered. As per 3.2 no reply is considered.

3.4 Pandemic data

The information associated to the pandemic is retrieved by the means of a specific python package called covid19dh [15]. It wraps a free API providing a direct connection to the unified dataset obtained by Guidotti et al. [16] through the aggregation of several respected sources such as World Health Organization, European Centre for Disease Prevention and Control, United States Center of Disease and Control, etc.

4 Data storage

Once the data are acquired through the specific python script, they are stored to be easily identifiable and reachable by the successive processes. Every dataset is saved in the same dedicated folder within a pickle file named according to its content. Specifically, a pickle file is the result of the serialisation, alternatively known as pickling, of a python object structure into a byte stream. In comparison with the other mainstream file format, namely Comma-Separated Values (CSV), it speeds up the saving and loading execution time while requiring less disk space. Nevertheless, it is neither human readable nor cross-platform and it has some security issues that suggest to manage pickle files only if the source is trusted. Additionally, a No-SQL cloud-based database on MongoDB is configured to retrieve the datasets described so far. In comparison with SQL databases, the non relational counterparts offer a more flexible schema wherein documents hosted in the same collection can have completely different structures increasing the easiness of data management. Besides, No-SQL databases are easily scalable following a scale-out strategy and can reach higher performance and reliability through sharding and replication.

5 Data preprocessing

Data preparation is usually the phase that requires more time in the entire process since the inference results are significantly driven by the strategies here adopted to elaborate the collected information. Albeit a detailed description of the preprocessing steps performed is subsequently provided for each specific dataset, it is possible to define a general guideline. In most of the cases, the data acquired are dirty. Namely, they can exhibit lacking values or attributes due by data unavailability, hardware and software problems (incomplete) as well as errors or outliers caused by mistakes committed throughout the data collection, insertion and transmission (noisy). In the first step, each dataset is cleaned by handling the missing values, errors and outliers. Then, once the discrepancies among the sources are solved, they are merged into a single consistent dataset. Finally, data can be normalised if the attributes develop on non equal ranges and reduced in the presence of high dimensional samples.

5.1 Stock data & Indexes

Initially, the stock dataset is obtained combining the data regarding the sole Microsoft company with the indexes related to the leading listed company in the market. Then, a new version of the dataset is created ruling out some columns, i.e. attributes. In particular, the split, close and dividend variables are excluded since already combined in the adjusted close pricing. The latter is therefore usually preferred to the mere closing price since it is more technically accurate in reflecting the true stock value. On the other hand, the majority of the open, high, low, S&P500, S&P100 and DJIA attributes can also be discarded since they are significantly correlated. Correlation between variables is often exploited during the prediction process and sometimes can also reveal the presence of a causal relationship. Nevertheless, when working with linear or logistic regression algorithms, if two or more independent variables are highly correlated, the models performances can be negatively affected by a multicollinearity problem that leads to misleading results. Albeit the conclusions reported on the attribute relationships are derived from the visualisation of scattered-plot and correlation matrices (Fig. 1), most of the charts are omitted due to space limitation.

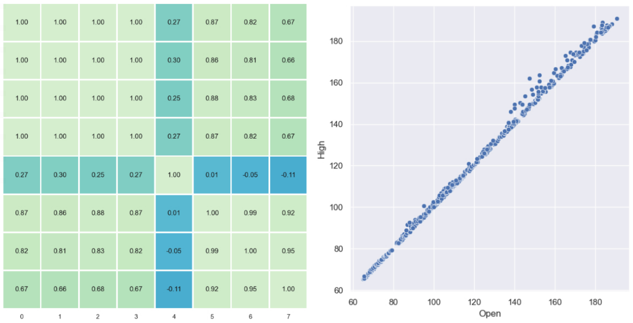

Fig. 1: The correlation matrix is on the left with the open, high, low, close, volume, S&P 100, S&P 500 and DJIA attributes that are represented from 0 to 7. On the right, a graph extracted from the scatter plot matrix.

The sole attribute retained together with the S&P 100 index and the adjusted closing price is the volume. Its magnitude over time indeed often allows to acquire a clearer perception of the strength behind increments and devaluations of the stock time-series. Then, for each attribute univariate outliers are detected by means of statistic methods, i.e. z-score and interquartile range (IQR) score, and graphical methodologies such as box-plots and scatter-plots. The written code also provides additional functions to find multivariate outliers through the deployment of a range of possible algorithms. Nonetheless, differently from outliers caused by incorrectly measured or entered data that should be dropped, it is always difficult to understand how to treat them when due to high data variance just like in this case. A common solution consists in testing the model with and without outliers in order to identify to which extent they influence the results and then acting accordingly. The initial strategy undertaken in this report intends to retain them. Finally, since stock data are measured under irregular intervals, i.e. exclusively during the business day, the dataset is re-sampled to include non-working days. All the missing not at random values originated by this last step are replaced using a linear interpolation to preserve patterns.

5.2 Tweets data

The majority of the 36 attributes defining the samples are dropped given their limited usefulness in the context explored. For example, the username or the conversation ids are not relevant to evaluate the overall public opinion. The only information retained for every post is the textual content, the date and the level of interaction generated, i.e. number of likes, replies and retweets. The tweets are cleaned in order to eliminate any type of hyper-textual link and superfluous space from the messages. Hashtags and cashtags are also discarded if not directly related to Microsoft. All the tweets published after the market closure, that cannot influence the daily session occurred, are modified to be considered as part of the successive day. Then, to extract the general attitude towards the company two different Natural Language Processing (NLP) models are investigated. Valence Aware Dictionary for sEntiment Reasoning (VADER) is a lexicon and rule-based sentiment analysis tool that is particularly designed to extract sentiment from social media platforms [17, 18]. It is able to evaluate slang words, emoticons, emojis, acronyms, word-shape and punctuation. After associating a negative or positive score to each word, it computes the probability that the entire text is negative, positive or neutral. The other solution adopted consists in a pre-trained supervised machine learning model provided by the Flair framework [19, 20]. It could lead to more accurate results but it is more complex and requires a larger amount of resources. Regardless the methodology used, the model outputs are combined to represent the tweet polarity with a value within the range [-1, 1] where 1 means highly positive and -1 extremely negative. Thereafter, two strategies are investigated. Whereas in the first the tweets are equally considered, in the second strategy the messages are weighted according to the number N and the significance W associated to likes l, replies rp and retweets rt. For each tweet the final score is obtained by means of the equation:

$$C = R \cdot \sum_{i} N_i \cdot W_i$$

where R is the result achieved by the selected sentiment analysis tool and $i \in \{l, rp, rt\}$. Note that the formula described is valid and computed only if the sum of the weights assigned to likes, replies and retweets is greater than 0. During the preprocessing of the dataset, no control is applied to detect the presence of duplicates since the interest is focused not only to the sentiment expressed but also to the reactions generated. In general, it could be reasonable to consider retweets more valuable with respect to likes since the level of the user exposure differs significantly. On the other hand, it is quite difficult to assign a proper weight to the replies since their number can be also correlated to a general sense of disagreement. Nonetheless, thanks to several experiments conducted it turns out that the former strategy could effectively lead to better results. Hence, the sentiment extracted for each day is achieved computing the mean of the results R of the tweets posted in the same 24 hours. As well as for the stock dataset the non-considered days, i.e. wherein no tweets are posted, are added. However, in this case the related missing values are substituted with 0. No interpolation or measure of centrality is used to replace lacking elements since they are originated by the fact that no post is published rather than by errors or else.

5.3 News data

The inclusion of the news dataset in the project has been challenging since the acquisition phase, characterised by a trade-off between the value and the number of the sources. Firstly, less Twitter accounts were scraped resulting into a modest number of observations that made the entire preprocessing system not enough reliable. Indeed, generally the relevance of the mistakes of the sentiment analysis tool increases with the lowering of the observations taken into account for every day. To overcome this problem, the selection was further extended, involving less followed and probably less authoritative Twitter users. The preprocessing applied to the news dataset follows exactly the same steps presented in the previous subsection. It might be worth observing that in this dataset the number of duplicates may be greater than the tweets dataset since it is usual for newspapers to tweet the same headlines several times. However, the strategy pursued is the same of that described in Section 5.2: no matter the source or the content of the post, the relevant information resides in the sentiment that can be extracted and shared by the public opinion.

5.4 Pandemic data

Initially, likewise the other datasets a set of variables not strictly necessary for the project aims are excluded retaining solely the columns that provide the number of confirmed, recovered and deaths. In general, the policies adopted by the countries to face the proliferation of the SARS-CoV-2 focus mainly, although not exclusively, on the number of new active cases, i.e. never affected people that contracted the virus. Therefore, after replacing all the missing values, wherein no cases are registered, with 0, the variables are combined to obtain the daily percentage change of the active cases.

5.5 Data integration

Before performing data integration, the existing discrepancies among the datasets are solved by uniforming how dates are expressed. The unified dataset obtained comprises 1124 daily observations defined by: the adjusted closing price and the volume of the Microsoft stocks; the S&P 100 index; the overall public sentiment extracted from both respected and common users on Twitter; and the percentage of new Covid19 cases. Albeit Section 5 focused exclusively on the observations within the pre-defined period, the pre-processing phase also involved the values of the exogenous variables measured within May 2020. Since the hyper-parameters of the models subsequently described in Section 7.1 are selected evaluating their performance on a validation set, any data normalization or reduction is performed after the dataset splitting. Obviously, when the dataset corresponds to a time-series wherein observations are not independent, it cannot be divided by picking random samples. The alternative adopted in this project to carry out the dataset segmentation is known as fixed partitioning and consists in separating the dataset into consecutive periods. Once the proper model hyper-parameters are selected, the training phase is re-executed on the union of the training and validation portions [21]. Finally, the independent variables are normalised inside the range $[0, 1]$ before being reduced by means of the Principal Component Analysis (PCA) algorithm and rescaled to develop on the same range of the closing price attribute.

6 Data exploration

This phase consists in exploring data in order to find relevant properties and anomalies, that can suggest the need of an additional pre-processing step, and identify salient patterns and relationships between variables allowing to anticipate the successive inference outcomes. Firstly, it is possible to point out that the dataset obtained before applying any normalization and reduction is composed by quantitative attributes whose statistic properties are summarised in Tab. 1.

Statistic	Close	Volume	S&P 100	Sentim.	Mood	Covid
Mean	108.25	2.87 E^{7}	1.22 E^{3}	0.17	- 0.13	0.02
Std	310.11	1.41 E^{7}	1.08 E^{2}	0.06	0.44	0.31
Min	61.26	7.42 E^{6}	1.03 E^{3}	- 0.06	- 0.99	0.00
25%	84.39	1.93 E^{7}	1.14 E^{3}	0.13	- 0.42	0.00
Median	103.84	2.45 E^{7}	1.21 E^{3}	0.17	- 0.01	0.00
75%	134.37	3.30 E^{7}	1.28 E^{3}	0.20	0.14	0.00
Max	186.72	1.11 E^{8}	1.51 E^{3}	0.44	0.99	10.00

Table 1: Statistical averages and variability of the dataset.

Specifically, the sentiment and mood columns explain the daily average of the outcomes returned by the sentiment analysis tools when applied to the tweets and news datasets respectively. From the information reported in the table above and in Fig. 2, it could seem that the covid attribute exhibits an unusual behaviour. Nonetheless, it is reasonably caused by the fact that the virus thrived only during 2020 and therefore most of the measured values are equal to 0.

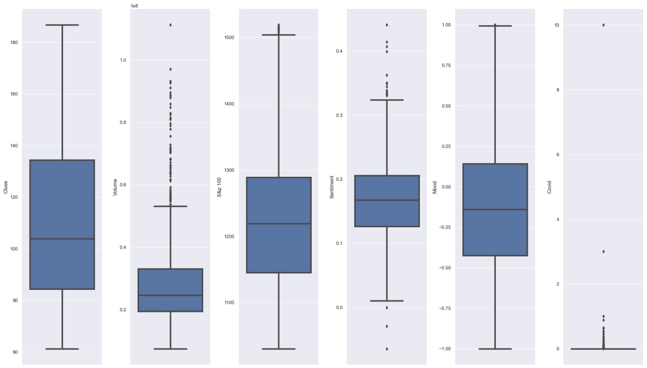

Fig. 2: Uni-variate boxplots of the close, volume, S&P100, Sentiment, Mood and Covid columns respectively.

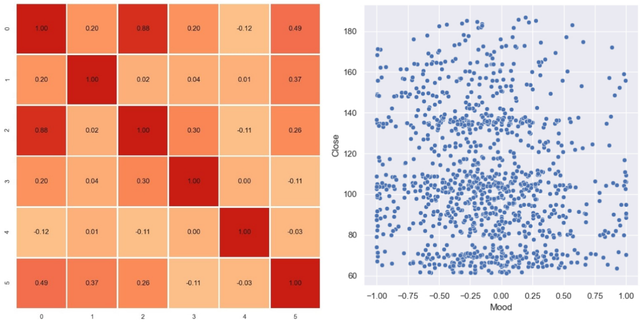

Fig. 3: The correlation matrix is on the left with the close, volume, S&P 100, sentiment, mood and covid attributes that are represented from 0 to 5. On the right, a graph extracted from the scatter plot matrix.

Additionally, existing relationships among variables are investigated exploring the scatter plot and correlation matrices of the dataset. In particular, it could be worth to analyse the connection between the mood and the close attributes. As visible from Fig. 3, the sentiment extracted from the news via the Flair pre-trained model is slightly negatively correlated to the closing price. Moreover, the strength of the correlation and the presented scatter plot put in evidence that the relationship is almost independent. Since the little information that the mood column brings appears to be paradoxical, i.e. the stock value increases when negative news about Microsoft are published, it may be necessary to rule out the variable before the prediction process. This is probably the result of a set of factors such as the limited number of daily data, the mistakes computed by the sentiment analysis tool or/and an imperfect selection of the sources during the acquisition phase. A time series is typically composed by several components, namely trend, seasonality, cyclicity and random noise. According to the type of the decomposition the time series can be rebuilt by adding or multiplying the isolated parts. In general, the multiplicative decomposition is more appropriate for economic time series, such as in this case, wherein the size of the fluctuations increases along with the magnitude of the variable. The auto-correlation function, that reports the relationships between delayed values of a time series, is often helpful to identify the presence of patterns. The existence of a trend in the data is proved by the auto-correlation function values that are greater for limited lags and that decrease gradually with larger delays. Seasonality occurs when the values of the function rise with a specific frequency. The correlogram in Fig. 4 shows the existence of a trend but no seasonal patterns.

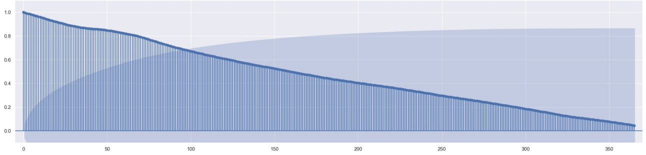

Fig. 4: Correlogram of the closing price time series.

Identical conclusions can be reached from the decomposition reported in Fig. 5 where the seasonal component is negligible. It moves in a infinitesimal neighbourhood around 1 not producing any relevant effect on the time series. The code also provides a function to generate scatter plot matrices where each data point can be coloured according to the year, the month, the day or the weekday in which it was measured. Although the plots are omitted for space limitations, the lack of patterns in the charts enhances the conducted arguments.

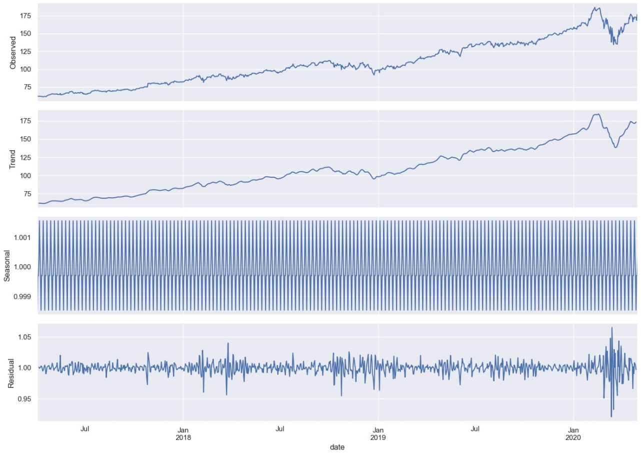

Fig. 5: Decomposition of the closing price time series.

Further information on the time series are extracted just by superimposing the averages and the standard deviation computed within a sliding window on the stock value plot (Fig. 6). The Simple Moving Average (SMA) and Exponential moving average (EMA) make clear the existence of a growing trend. The difference between the two is how the observations considered, i.e. within the rolling window, are weighted. The latter is more dynamic since the most recent values have a greater influence. On the other hand, the computation of the standard deviation over time enables to understand the volatility magnitude of the time series. As visible from Fig. 6, it increases especially in the final period taken into account explaining in part the characteristics of the residuals plot in Fig. 5. Besides, the image above (Fig. 6) highlights that the series is non stationary, i.e. neither the variance nor the mean are invariant. Since many forecasting models require at least trend stationarity, a function is provided to find the number of times a series needs to be differenced to reject with a high level of confidence the null hypothesis (H0) of the Augmented Dickey-Fuller Test (ADF). In this case, a sole order of differencing turns out to be enough. Moving averages are mainstream lagging indicators able to detect support and resistance areas whereby the time series is above or below the moving average line. Based on this, a trader might buy shares when the series exceeds the moving average, such as in the last days represented in Fig. 6, and sell otherwise. Finally, from the charts proposed it is possible to observe that in general Microsoft stock rises progressively and with consistence except for sporadic times. The only significant drop, although followed by a rapid resumption, occurred in March 2020 with the fast proliferation of the coronavirus SARS-CoV-2.

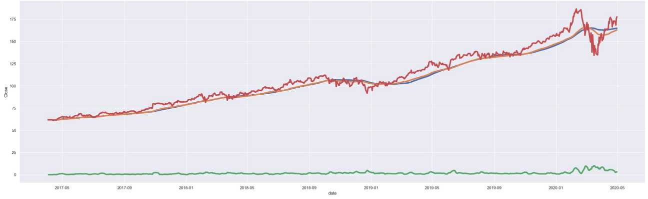

Fig. 6: The figure depicts the SMA (blue) and EMA (orange) along with the adjusted closing price (red). The green line at the bottom represents the standard deviation computed on the same sliding window of length 90, i.e. a quarter.

Hypothesis testing relies on the formulation of a null and of an alternative hypothesis (H0 and H1) concerning a specific property of a population. Once the test typology is selected and the test performed, H0 can be rejected in case the result achieved is lower than a pre-defined significance level [22]. Otherwise, there may be not enough evidence to support a conclusion. The majority of the tests conducted to evaluate the relationships between attributes take into account the extent to which a range of values of a variable drives the daily change of an other variable. For instance, it comes up, through the execution of a two-samples one-tail T-test, that the average of the daily changes of the stock closing price related to the 35% of the volume highest values is higher than the average of the remaining values. The results of a different two-tails T-test exhibits that in the days where the Covid-19 active cases are almost doubled the volume daily variation reports a distinct behaviour. A number of tests is also performed throughout the inference phase to better understand the properties of the models residuals.

7 Data inference

In this Section, two distinct models to predict values from the Close and Volume columns of the stock dataset are compared. The one reporting best results is then selected to perform the same task but leveraging also the additional exogenous variables.

7.1 Models

The first methodology investigated belongs to the family of the Auto Regressive Integrated Moving Average (ARIMA) models. As well as the ARMA model, it exploits a linear combination of the previous values of the variable of interest (auto regression) along with past forecast errors (moving average) to predict the future values of the time series y_t:

$$\phi(L)(y_t - X_t\beta) = \theta(L)\epsilon_t$$

where $X(t)\beta$ is the exogenous variables effect, ϵ_t the forecasting errors and L the back-shift operator, alternatively known as lag operator. In a more readable extended version the equation, with $u_t = y_t$ - $X_t\beta$, becomes:

$$\begin{aligned} y_t = X_t\beta + \phi_1 u_t + ... + \phi_p u_{t-p} + \\ + \epsilon_t - \theta_1\epsilon_{t-1} - ... - \theta_q\epsilon_{t-q} \end{aligned}$$

However, as alredy anticipated, a mainstream assumption in several forecasting techniques is stationarity, i.e. statistical properties do not vary over time. ARIMA is able to render trend stationary a non stationary time series computing the differences among a number of consecutive observations (differencing) [23]. In the previous formula, $\phi(L)$ is hence replaced by $\nabla^d\phi(L)$ where d is the differencing order and ∇ = (1 - L). On the other hand, Facebook Prophet is a modern regression model that stands out for its robustness to outliers, missing data and significant time series fluctuations [24, 25]. Furthermore, it does not require stationary ingress. The model is the result of the combination of different components:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t + X(t)\beta$$

where $g(t), s(t), h(t), \varepsilon_t$ represent respectively the portions associated to non-periodic and periodic changes, holidays effect and errors caused by unusual behaviours not captured by the model.

7.2 Metrics

In addition to arguments derived from residuals characteristics the model is selected using a series of well-known metrics. Root Mean Squared Errors (RMSE) enables to weight bigger errors and avoid the deletion between negative and positive errors thanks to the squared elevation. Mean Absolute Percentage Error (MAPE) operates on the error absolute values again not allowing the compensation of positive and negative differences. Moreover, differently from RMSE and Mean Absolute Error (MAE), it is scale independent. Albeit it is probably the most used metric in business predictions, it can mess with observations equal to 0 and it is also asymmetric penalising more the over-forecasting errors. On the other hand, in the Mean Percentage Error (MPE) the negative and positive errors can offset each other. For this reason, the formula is often adopted as a measure of the bias introduced throughout the foresight. Finally, the correlation between the true and predicted values is also reported.

7.3 Model selection

The models comparison is based on the ARIMA and Facebook Prophet results achieved starting from solely the company's data, i.e. the Volume and the Close columns of the ultimate dataset. For ARIMA, a specific function provided by the pmdarima package [26] allowed during the validation phase to search the optimal values of the majority of the hyper parameters according to the lower Akaike Information Criterion (AIC) score obtained. In addition to the metrics, the choice of the model is driven by the residuals analysis. Residuals are the part of information that the model has not captured in the data [23]. In case they are correlated or have a non zero mean it means respectively that some information useful for the prediction is left over or that the foresight is biased. Residuals with constant variance and that are normally distributed could also be required, even if not mandatory, to facilitate the computation of the prediction intervals. ARIMA is the model selected to perform the prediction exploiting the entire final dataset. The results reported in Tab. 2 exhibit that it exceeds Facebook Prophet performance in the majority of the metrics taken into account. Moreover, its residuals, as noticeable in Fig. 7, seem to behave better. Although both the residuals of the models have a mean that is significantly close to 0, Facebook Prophet's residuals follow a distribution that is nearer to a normal distribution. It may be plausible that this is caused by the fact that the Facebook Prophet model is more capable of capturing the volatility of the time series, especially in the last days where its magnitude increases. In any case, as previously mentioned, it is firstly required that adjacent residuals are not correlated with each other. The auto-correlation plot shown within the second row of Fig. 7 demonstrates that ARIMA respects this requirement.

Model	RMSE	MAE	MAPE	MPE	Corr
Arima	3.048	2.522	0.014	-0.009	0.508
Prophet	4.912	4.454	0.024	-0.024	0.605
Arima	3.680	3.257	0.018	-0.009	0.600

Table 2: The results achieved by both the models starting from the Close and Volume columns are displayed at the top, whereas the results obtained by the selected model starting from the entire dataset collected are at the bottom.

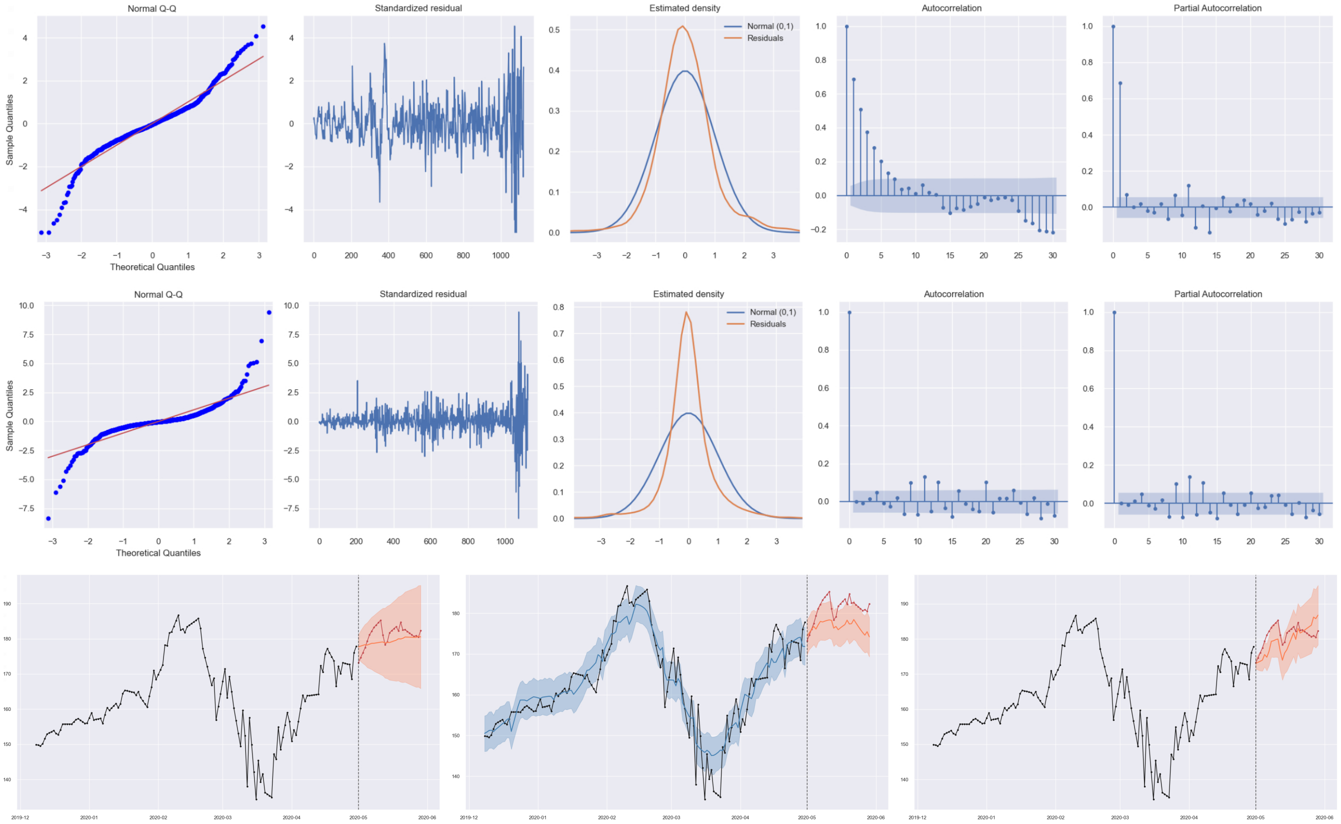

Fig. 7: In the first two rows, a visualisation of the residuals properties of the Facebook Prophet and ARIMA models is reported. At the bottom, the results achieved with both the models (ARIMA and Prophet) starting from solely stock data are displayed along with the prediction made by the selected model (ARIMA) on the entire dataset collected.

7.4 Results analysis

The results discussed are achieved by the chosen model that forecasts the closing price of the Microsoft stock leveraging further external data in addition to the measured stock price and volume values. Specifically, it also considers the effects that may be explained by the records of the ongoing pandemic, the behaviour of the 100 major listed companies as well as the sentiment extracted from the Twitter through the VADER sentiment tool. From an initial comparison (Tab. 2) between the outcomes of the same ARIMA model obtained before with the sole company's stock data and after with the entire dataset, the supplement of information seems to slightly decrease the model performance. However, the level of uncertainty of the model during the prediction of the new data point values must be considered in the comparative. From the 1^{st} and the 3^{rd} pictures within the last row of Fig. 7, it is possible to observe that the prediction bands obtained when using the whole dataset are much less wider. Moreover, their width increases significantly slower than the prediction bands width of the model whose prediction is based only on the company's stock data. Furthermore, whereas in the first picture the predicted part of the time series follows a quite linear evolution, in the third image the foresight evolution seems to better track the variations of the true closing price values.

8 Conclusion

In this report, Microsoft stock closing prices of May 2020 are forecasted since data ranging from April 2017 to April 2020. The prediction is performed firstly using solely company's stock data and subsequently exploiting supplement information about the public mood on Twitter, others market indexes and the current pandemic. Albeit good performance are reached directly from the close and volume stock attributes, the results can be enhanced through the addition of further information especially in terms of the model's level of uncertainty. Undoubtedly, onward improvements can for instance be achieved extending the tweets and the news datasets with posts published in other languages. This, together with the adoption of a more performing and tailored sentiment analysis tool, may allow to overcome the main difficulties encountered concerning the news dataset. Finally, a different approach that could lead to better performance consists in handling the time series and its volatility separately by means of the combination of the ARIMA and GARCH models.

9 References

[1] Paul H Cootner, *The random character of stock market prices*, MIT press, 1967.

[2] Eugene F. Fama, "Efficient capital markets: Ii," *The journal of finance*, vol. 46, no. 5, pp. 1575–1617, 1991.

[3] Thomas Hellström and Kenneth Holmström, "Predicting the stock market," 1998.

[4] Manolis G. Kavussanos and Everton Dockery, "A multivariate test for stock market efficiency: the case of ase," *Applied Financial Economics*, vol. 11, no. 5, pp. 573–579, 2001.

[5] Bo Qian and Khaled Rasheed, "Stock market prediction with multiple classifiers," *Applied Intelligence*, vol. 26, no. 1, pp. 25–33, 2007.

[6] Robert P. Schumaker and Hsinchun Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 2, pp. 1–19, 2009.

[7] Johan Bollen, Huina Mao, and Xiaojun Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[8] Anshul Mittal and Arpit Goel, "Stock prediction using twitter sentiment analysis," *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, vol. 15, 2012.

[9] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi, "Sentiment analysis of twitter data for predicting stock market movements," in *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE, 2016, pp. 1345–1350.

[10] Mieszko Mazur, Man Dang, and Miguel Vega, "Covid-19 and the march 2020 stock market crash. evidence from s&p1500," *Finance Research Letters*, p. 101690, 2020.

[11] "Alpha vantage documentation," https://www.alphavantage.co/documentation, Accessed: 29 December, 2020.

[12] "Alpha vantage python package," https://github.com/RomelTorres/alpha_vantage/tree/master, Accessed: 29 December, 2020.

[13] "Pandas data reader," https://github.com/pydata/pandas-datareader, Accessed: 28 December, 2020.

[14] "Twint package," https://github.com/twintproject/twint, Accessed: 30 December, 2020.

[15] "Covid19 data hub," https://github.com/covid19datahub/COVID19, Accessed: 30 December, 2020.

[16] Emanuele Guidotti and David Ardia, "Covid-19 data hub," *Journal of Open Source Software*, vol. 5, no. 51, pp. 2376, 2020.

[17] Che Gilbert and Erric Hutto, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14).*, 2014, vol. 81, p. 82.

[18] "Vader (valence aware dictionary and sentiment reasoner)," https://github.com/cjhutto/vaderSentiment, Accessed: 31 December, 2020.

[19] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[20] "Flair," https://github.com/flairNLP/flair, Accessed: 31 December, 2020.

[21] Laurence Moroney, "Sequences, time series and prediction," https://www.coursera.org/learn/tensorflow-sequences-time-series-and-prediction, Accessed: 25 December, 2020.

[22] Ivan Svetunkov, "Time series analysis and forecasting with adam: Lancaster, uk.," openforecast.org/adam, Accessed: 31 December, 2020.

[23] Rob J. Hyndman and George Athanasopoulos, *Forecasting: principles and practice*, OTexts, 2018.

[24] Sean J. Taylor and Benjamin Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[25] Sean J. Taylor and Benjamin Letham, "Prophet: Automatic forecasting procedure," https://github.com/facebook/prophet, Accessed: 31 December, 2020.

[26] "Pyramid arima," https://pypi.org/project/pmdarima/, Accessed: 30 December, 2020.