Master's Degree in Data Science

Investigation of Sentiment Importance on Intraday Stock Returns

Authors: Michele Costa, Alessandro De Sanctis, Laurits Marschall, Seyed Hamed Mirsadeghi

Directors: Ioannis Arapakis, Carlos Segura Perales

June 2018

ABSTRACT IN ENGLISH

The main goal of our Master Project is to predict intraday stock market movements using two different kinds of input features: financial indicators and sentiments from news and tweets. While the former are part of the common technical analysis of financial econometric models, the extracted sentiment of news articles and tweets from Twitters are also proven to correlate with stock markets movements. Our paper aims at contributing to the existing academic and professional knowledge in two main directions. First, we evaluate three different approaches to extract the sentiment from both social and mass media based on its forecasting power. Second, we deploy a battery of engineered features based on the sentiment, together with the financial indicators, in a machine learning model for a fine-grained minute-level forecasting exercise. In the end, two different classes of models are fitted to test the forecasting power of the combined input features. We estimated a classical ARIMA-model, and an XGBoost-model as machine learning algorithm. We collected data on the companies Apple, JPMorgan Chase, Exxon Mobil, and Boeing.

ABSTRACT IN CATALAN

L'objectiu principal del nostre Projecte de Màster és predir els moviments intradia del mercat de valors utilitzant dos tipus diferents de característiques d'entrada: indicadors financers i sentiments de notícies i piulades. Mentre els primers són part de l'anàlisi tècnica comú dels models economètrics financers, els sentiments extrets d'articles de notícies i piulades de twittaires també tenen una correlació demostrada amb els moviments del mercat de valors. El nostre article vol contribuir al coneixement acadèmic i professional en dues direccions principals. En primer lloc, avaluem tres aproximacions diferents per extreure els sentiments de les xarxes socials i els mitjans de masses basant-se en els seus poders de predicció. En segon lloc, despleguem una bateria de característiques d'enginyeria basades en el sentiment, juntament amb indicadors financers, en un model d'aprenentatge automàtic per a un exercici de predicció desgranat al minut. Finalment, es fan dues classes diferents de models per testejar el poder de predicció de les característiques d'entrada combinades. Hem estimat un model ARIMA clàssic i un model XGBoost com a algoritme d'aprenentatge automàtic. Hem recavat dades de les companyies Apple, JPMorgan Chase, Exxon Mobil i Boeing.

KEYWORDS IN ENGLISH: Machine Learning, Sentiment Analysis, XGBoost, Finance

KEYWORDS IN CATALAN: Aprenentatge automàtic, Anàlisi de sentiments, XGBoost, Finances

Barcelona Graduate School of Economics

M.Sc. Data Science

Investigation of Sentiment Importance on Intraday Stock Returns

Masters' Project

A.Y. 2017 - 2018

by

Michele Costa, Laurits Marschall,
Hamed Mirsadeghi, Alessandro De Sanctis

Supervisors

Ioannis Arapakis, Carlos Segura Perales

Contents

1 Introduction

The main goal of this Master Project is to predict intraday stock market movements using two different kinds of input features. Financial indicators are used similar to a classical financial econometric model and, on the other hand, the extracted sentiment of news articles and Twitter tweets is also used as input features. It is proven that the sentiment of news articles and tweets correlates to stock market movements (Bordino and al. (2014)).

This paper aims at contributing to the existing academic and professional knowledge in two main directions. On one side, the sentiment of both social and mass media is used as an input feature of prediction models to validate its forecasting power. Three different approaches are taken to extract the sentiment of news and tweets and evaluated separately. The second novelty consists on the deployment of these features, together with the technical financial indicators, within a machine learning model for a fine-grained minute-level forecasting exercise. In the end, two different classes of models are fitted to test the forecasting power of the combined input features. We estimated a classical ARIMA-model, and a XGBoost-model as machine learning algorithm.

The analysis is constrained to the companies Apple, JPMorgan Chase, Exxon Mobil, and Boeing.

2 Literature Review

The interest in stock markets, and the corresponding return forecasting exercise, is long-dated and different academic disciplines, such as finance, economics, statistics, and computer science, have experimented different approaches. The recent developments in computing power, distributed system, and data sources has contributed to a renewed effort and innovative approaches.

Some work has been carried out by integrating unstructured information in stock prediction models. Borrowing the specific technique from other fields, such as epidemics, Bordino and al. (2012) successfully combined web search engine activity with stock market movements. In fact, they found a statistically significant relationship by looking at the correlation between Nasdaq companies' daily trading volumes and *Yahoo! Search* activity on the specific stock. Browsing activity also provides relevant forecasting power to predict trading volumes a few days in advance. It is worth mentioning that this research provides support for the idea of *wisdom of the crowd*, where the aggregation of idiosyncratic and un-coordinated user behaviour results in a powerful prediction tool. The following publication by Bordino and al. (2014) in 2014 took the research one step ahead, linking web search volumes and investors sentiment. Furthermore, the analysis is extended to 2600 traded stocks, exploiting sector-level and industry-level clustering, and it is conducted both at daily and intraday levels.

Another important component of the project is the sentiment analysis carried out on different sources and document types. Nassirtoussi and al. (2014) dived into the interdisciplinary nature of textual analysis for finance purposes, and outlined a more defined shape of the theoretical and empirical base of this emerging branch. They observed that, although no common accepted methodology exists, most of the publication shares some feature selection and representation

strategies, along the machine learning model involvement and the evaluation mechanism. From the text analysis point of view, interesting publications are also those by Lin and He (2009), and Joshi et al. (2016). The two papers implemented sentiment analysis combined with other NLP techniques in order to explore the data set in multiple directions. The first one presents a joint sentiment and topic (JST) modelling algorithm that can be applied in unsupervised setting. They recognized that previous work attempted to obtain the results in two-steps procedure, and that results were suffering from lack of informative content as well as lack of detection of domain-specific terminology.

The second publication developed a similar approach to what is shown in this project, as it classifies 60 millions tweets based on their sentiment score. The authors develop the sentiment language model using a data set of 200'000 labelled product reviews, and they adopt Named Entity Recognition methodology to increase the accuracy of the sentiment analysis. Remaining in the domain of sentiment analysis for the financial industry, the survey by Kerney and Liu (2014) identifies three different sources of textual information to be analyzed: corporate filings, media-based, and internet-based documents. It exists a trade-off between degree of relevance and frequency of observations, as the board documents and quarterly disclosures are often more informative. This integrates well with the approach presented later in the Sentiment Analysis section, for Loughran-McDonald dictionaries. The authors point to relevant contribution by Sinha (2010) and Ferguson et al. (2015) who structured their work on substantial data-sets of more than 200'000 articles. Chen et al. (2017) explored several methodology to extract the sentiment analysis from a Twitter data set, labelled by experts as well as by two span determination algorithm. These observations are combined with the articles eventually referenced in the tweets. The authors test 15 sentiment dictionaries both of real value and binary nature. Sul et al. (2017) also kept the focus on Twitter, with a data-set of 2.5 millions entries regarding the SP500 index companies, and tries to identify whether the cumulative sentiment for each entity is relevant for forecasting the stock market performance. The authors attempts several time horizon prediction, from daily to 20 days ahead, and they notice an interesting pattern in the nature of the accounts providing the highest relevance for the possible trading strategy. Indeed, the sentiment gathered from tweets of users with the number of followers below the median shows forecasting power on 10 and 20-day time frame, and the derived trading strategy yields meaningful annualized returns in the order of 11-15 percent.

Lee and al. (2014) restricted the attention to 8-K filings by USA listed companies, in order to process insightful information directly from the corporations without reducing the observation frequency, as the companies are obliged to publish these documents every time that a meaningful business event takes place. The forecasting exercise is limited to the time periods following the publication of such reports. They showed that the prediction performance highly benefits from the inclusion of textual features along pure financial data, compared to the benchmark model with financial information only.

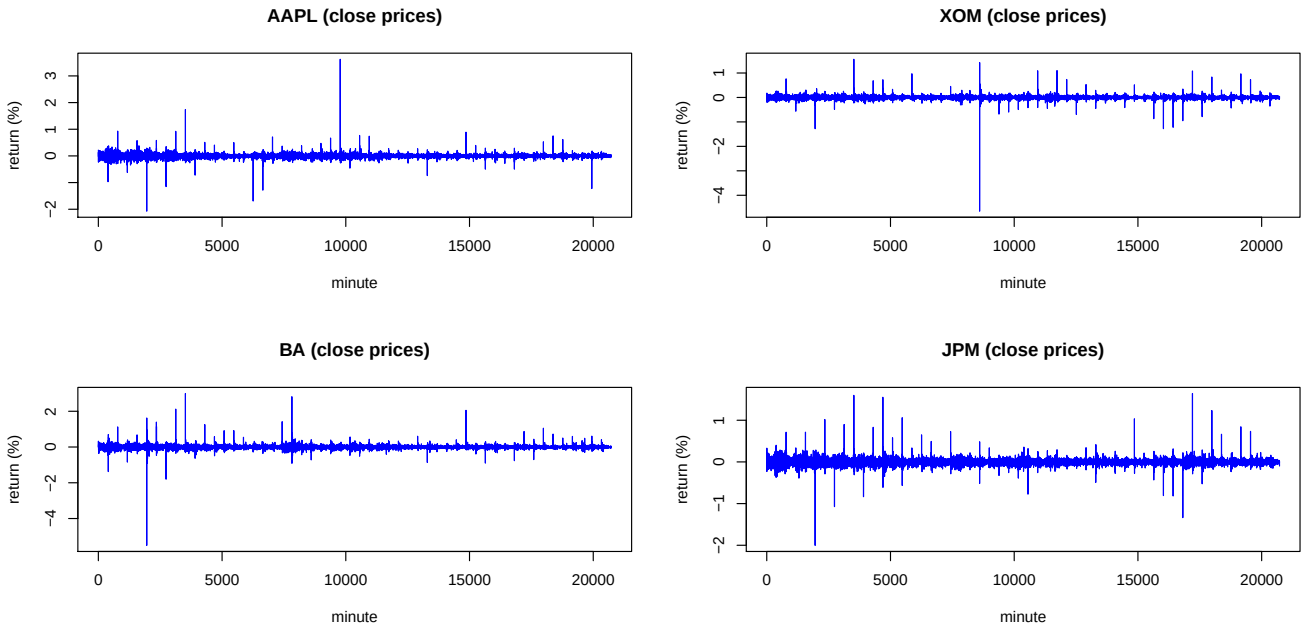
A recent contribution attempting to bridge the gap between classical statistics and Machine Learning models within the forecasting sphere is the one by Makridakis et al. (2018), that evaluated the performance of eight popular Machine learning models and compared them with eight classical ones. Both accuracy and computational complexity results favor the statistical approach, that benefits from lower model complexity and the direct possibility to apply optimization pro-

cesses for parameters and models choice. The authors suggested to lay out a standard and accepted methodology to objectively assess the quality of forecasting models through large scale competitions, to improve the understanding of the mechanisms behind the prediction generation and to enrich these results with considerations about the uncertainty.

3 Experimental Methodology

3.1 Data Collection

The analysis is based on prices and volume of the four aforementioned companies of the Dow Jones Industrial Average, one of the most important stock market indexes in the world. Companies in this index are publicly owned entities, based in the Unites States and characterized by a large market capitalization. Intraday observations at minute level are collected, by using the free data API from Alpha Vantage[1]. For each of the four stocks, the database includes open, low, high, close prices and volumes of the trades of every minute from April 1st. The figure below shows the minute-by-minute return (%) of the four stocks we focused on (AAPL for Apple, XOM for Exxon Mobil, BA for Boeing and JPM for JPMorgan Chase).



We collected the data from the Alpha Vantage API on a regular basis, followed by some consistency checks to detect potentially missing observations in the time series. This problem affected only between 0 and 2% of the observations for each stock. We imputed the few missing values by *backward filling*. The next step was to compute a variety of technical indicators to take into account possible trends in the price movements and their persistence.

[1] https://www.alphavantage.co/

3.1.1 Financial Indicators

This section presents the different technical indicators used in the prediction models. All indicators were computed using the free Technical Analysis Library[2]. Technical indicators are functions of prices and volumes of trades. As described in Nassirtoussi and al. (2014), these financial indicators are constructed and used by technical analysts with the intention to detect recurrent patterns in prices. The authors claim that even if technical indicators often lack predictive power, they are still widely spread among market brokers and participants.

We computed eleven technical indicators chosen to be the most used by the community. They can be split in moving average rules, relative strength rules, and trading range breakout rules. Their definitions come from FM Labs, a computer consulting company specialized in financial applications[3].

- SMA

  Moving Averages are used to smooth consecutive data points to help eliminate noise and identify trends. Each output value is the average of the previous n values. In a Simple Moving Average, each value in the time period carries equal weight, and values outside of the time period are not included in the average. This makes it less responsive to recent changes in the data, which can be useful for filtering out those changes.

- EMA

  The Exponential Moving Average is a staple of technical analysis and is used in countless technical indicators. In a Simple Moving Average, each value in the time period carries equal weight, and values outside of the time period are not included in the average. However, the Exponential Moving Average is a cumulative calculation, including all data. Past values have a diminishing contribution to the average, while more recent values have a greater contribution. This method allows the moving average to be more responsive to changes in the data.

- MACD

  The Moving Average Convergence Divergence (MACD) is the difference between two Exponential Moving Averages. The Signal line is an Exponential Moving Average of the MACD. The MACD signals trend changes and indicates the start of new trend direction. High values indicate overbought conditions, low values indicate oversold conditions. Divergence with the price indicates an end to the current trend, especially if the MACD is at extreme high or low values. When the MACD line crosses above the signal line a buy signal is generated. When the MACD crosses below the signal line a sell signal is generated. To confirm the signal, the MACD should be above zero for a buy, and below zero for a sell.

- STOCH

  The Stochastic Oscillator measures where the close is in relation to the recent trading range. The values range from zero to 100. %D values over 75 indicate an overbought condition; values under 25 indicate an oversold condition. When the Fast %D crosses above the Slow

[2]TA-Lib API (http://ta-lib.org/)

[3]https://www.fmlabs.com/reference/default.htm

%D, it is a buy signal; when it crosses below, it is a sell signal. The Raw %K is generally considered too erratic to use for crossover signals.

- RSI

  The Relative Strength Index (RSI) calculates a ratio of the recent upward price movements to the absolute price movement. The RSI ranges from 0 to 100. The RSI is interpreted as an overbought/oversold indicator when the value is over 70/below 30. You can also look for divergence with price. If the price is making new highs/lows, and the RSI is not, it indicates a reversal.

- ADX

  The ADX is a Welles Wilder style moving average of the Directional Movement Index (DX). The values range from 0 to 100, but rarely get above 60. To interpret the ADX, consider a high number to be a strong trend, and a low number, a weak trend. The direction of the ADX line is important for reading trend strength. When the ADX line is rising, trend strength is increasing and price moves in the direction of the trend. When the line is falling, trend strength is decreasing, and price enters a period of retracement or consolidation.

- CCI

  The CCI is designed to detect beginning and ending market trends. The range of 100 to -100 is the normal trading range. CCI values outside of this range indicate overbought or oversold conditions. You can also look for price divergence in the CCI. If the price is making new highs, and the CCI is not, then a price correction is likely.

- AROON

  The word *aroon* is Sanskrit for "dawn's early light." The Aroon indicator attempts to show when a new trend is dawning. The indicator consists of two lines (Up and Down) that measure how long it has been since the highest high/lowest low has occurred within an n period range. When the Aroon Up is staying between 70 and 100 then it indicates an upward trend. When the Aroon Down is staying between 70 and 100 then it indicates an downward trend. A strong upward trend is indicated when the Aroon Up is above 70 while the Aroon Down is below 30. Likewise, a strong downward trend is indicated when the Aroon Down is above 70 while the Aroon Up is below 30. Also look for crossovers. When the Aroon Down crosses above the Aroon Up, it indicates a weakening of the upward trend (and viceversa).

- BBANDS

  Bollinger Bands consist of three lines. The middle band is a simple moving average (generally 20 periods) of the typical price (TP). The upper and lower bands are F standard deviations (generally 2) above and below the middle band. The bands widen and narrow when the volatility of the price is higher or lower, respectively. Bollinger Bands do not, in themselves, generate buy or sell signals; they are an indicator of overbought or oversold conditions. When the price is near the upper or lower band it indicates that a reversal may be imminent. The middle band becomes a support or resistance level. The upper and lower bands can also be interpreted as price targets. When the price bounces off of the

lower band and crosses the middle band, then the upper band becomes the price target.

- AD

  The Accumulation/Distribution Line is similar to the On Balance Volume (OBV), which sums the volume times +1/-1 based on whether the close is higher than the previous close. The Accumulation/Distribution indicator, however multiplies the volume by the close location value (CLV). The CLV is based on the movement of the issue within a single bar and can be +1, -1 or zero. The Accumulation/Distribution Line is interpreted by looking for a divergence in the direction of the indicator relative to price. If the Accumulation/Distribution Line is trending upward it indicates that the price may follow. Also, if the Accumulation/Distribution Line becomes flat while the price is still rising (or falling) then it signals an impending flattening of the price.

- OBV

  The On Balance Volume (OBV) is a cumulative total of the up and down volume. When the close is higher than the previous close, the volume is added to the running total, and when the close is lower than the previous close, the volume is subtracted from the running total. To interpret the OBV, look for the OBV to move with the price or precede price moves. If the price moves before the OBV, then it is a non-confirmed move. A series of rising peaks, or falling troughs, in the OBV indicates a strong trend. If the OBV is flat, then the market is not trending.

3.1.2 Twitter

As a social network, Twitter provides sentiment information on a high frequency regarding mentioned topics or companies. Tweets were selected by simple search-queries of the four respective company names.

Due to the amount of data collection necessary to extrapolate a meaningful result from the Twitter analysis, two different approaches have been used. Firstly, the official twitter API was used to download recent tweets and constantly update the databases during the time span of the Master project. The API provides tweets in our case based on search term queries of the actual text and hash-tags.

The API, however, only allows free-tier users to track nine days back in time to search for specific tweets and puts a restrictive limit on the number of downloaded tweets. Therefore, in a second approach the tweets were scraped directly from the web in order to go further back in time and circumvent quantity limits of the API. It means that all the information regarding the tweets is directly extracted from the HTML-code. Unfortunately, the second approach is, therefore, not capable of retrieving further information regarding the user behind the respective post despite its identification number. The downsides of the scraping approach were its very slow speed and the fact that Twitter blocked our the requests repeatedly.

Both data streams were consolidated and pre-processed to strip the text from unnecessary items. Only English tweets were collected for sentiment analysis which were in fact the vast majority of the tweets regarding the selected Dow Jones companies. Private and business twitter accounts were not treated differently throughout the whole analysis. However, tweets were weighted in a

specific way by taking into account the number of *retweets* of the respective tweet to account for the potential reach and importance of the tweet.

The number of available tweets differs vastly for the four selected companies of the Dow Jones index. Most of these differences can be easily explained by the differing size or business sector of the companies. By a vast margin, Apple is the company most mentioned or talked about on Twitter. Below the number of tweets per day for each stock are plotted. The numbers differ largely within and across stocks. However, the partially drastic spikes can be easily related to real-time events. It turned out that the Twitter API, in the end, provided less data than the scraping approach. The very unbalanced data set in terms of available tweets per day might be an issue later on for prediction reasons but could not be avoided considering the limitations of both approaches to collect tweets.

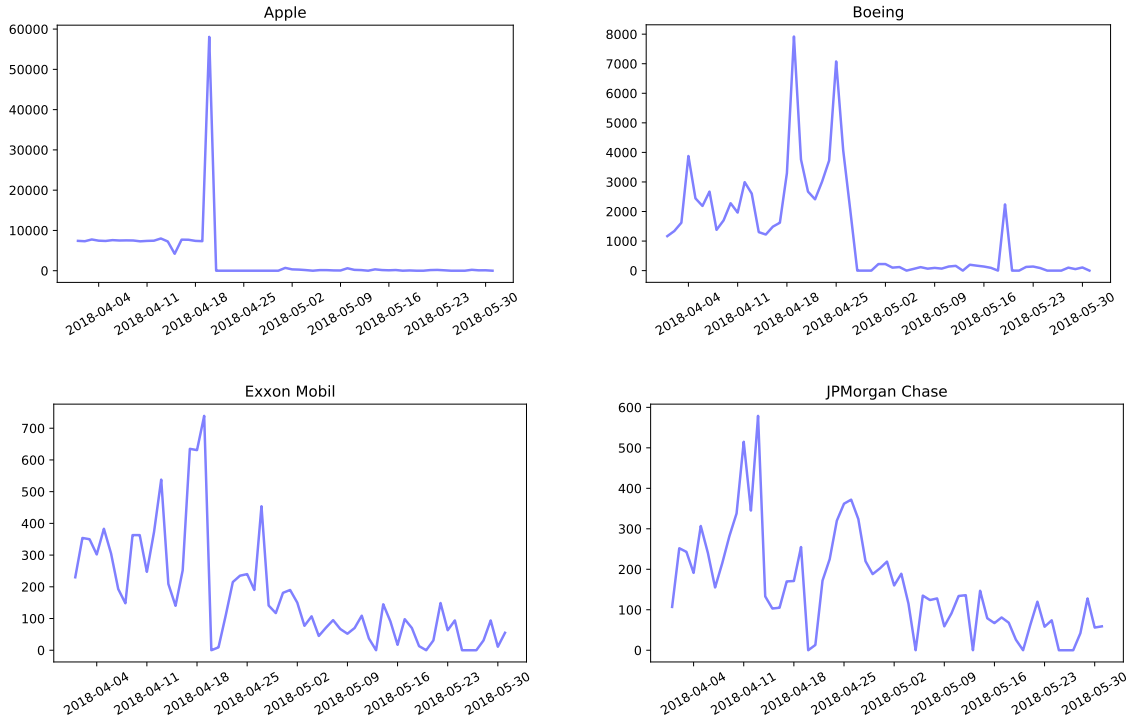


Figure 1: Tweets per company over time

3.1.3 Online News

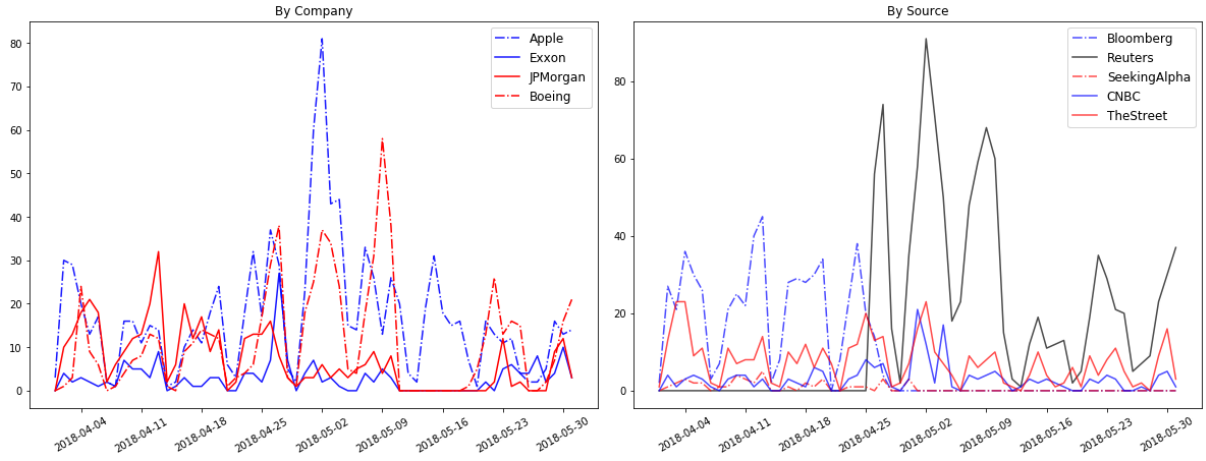
A variety of news sources and types are used to capture the wide sentiment about Dow Jones-listed companies by specialized audience, as well as by the general public. The choice of a wider pool of news outlets and sources also reduces possible reporting biases unrepresentative of the overall market.

The analysis is carried out on news outlets and web portals that have a major focus on finance. The first three of the lists are recognized to be among the most relevant business sources alongside Wall Street Journal, New York Times, Financial Times and The Economist.

- bloomberg.com : Bloomberg News is an international news agency whose content is disseminated through different channels and read by financial professionals thanks to the Bloomberg Terminals diffusion.

- cnbc.com : CNBC is an American basic cable, internet and satellite business news television channel, primarily carrying business day coverage of U.S. and international financial markets. The outlet is also present in different regions and countries, with a localised verion.
- reuters.com : Reuters is a news agency with global reach, recognized as one of the most largest provider of foreign news alongside Associated Press, and Agence France Press.
- marketwatch.com : MarketWatch operates a website providing business news, along analysis and market data. The outlet is part of the wider network of Dow Jones Media Group. It has hit a record 33.3 million unique visitors in February of 2018, and has an average of more than 16 million readers every month.[4]
- thestreet.com : TheStreet Inc. is an American financial news and services website founded by Jim Cramer, former hedge-fund manager and host of CNBC channel.
- seekingalpha.com : Seeking Alpha is a crowd-sourced content service for financial markets. The peculiarity of the news is that investors and industry experts contribute directly to the content, covering a broad range of stocks, asset classes, ETF and investment strategies. Seeking Alpha reportedly has an active community of 13 millions readers [5] .

The collected data includes articles specifically focusing on the entities of interest. This is implemented with a systematic retrieval of the online published content, that is stored together with the corresponding meta data: title, news source, author, keywords, and tags. This procedure allows the monitoring of pure financial news, such as earnings publications and management statements, along expert opinions by journalists and investment specialists. Finally, some documents provide relevant views about legislation and intervention by public authorities, which may affect the business performance at single-entity level.



All news are weighted by an ad-hoc relevance metrics, such that the final outcome presents the stream of sentiment collection in a coherent manner. The document-specific sentiment metrics are weighted with source-level relevance measure as well as document-specific importance checks.

[4]https://www.marketwatch.com/companyinfo
[5]seekingalpha.com/page/who$_r eads_s a$

Three distinct methodologies are used for assessing the news' relevance. To begin with, it is recorded how often the entity name, in various formats, appears in the title and in the body. Beside this strict matching, the second technique incorporates also the list of correlated terms and companies that appear for the Google Engine searches. This provides a larger base to evaluate the importance of the respective article and allows to incorporate sentiment related to the entity of interest expressed through the comparison with other companies or similar service providers [6] [7]. Finally, the popularity of the news source/outlet is represented via a proxy of Google Trends data. The daily frequency of web searches for each news outlet is stored in the database and then combined with the sentiment analysis. It is worth mentioning that Google pre-processes the raw number of search engine queries by keyword, by standardizing the output by location (worldwide in this case) and time frame (90 days): the highest number of searches in any day is set equal to 100 and the rest of the series is re-scaled accordingly. On top of the keyword-specific processing, the data used in the analysis is expressed in relative terms to the most popular website (CNN), so that the idiosyncratic elements of each news portal do not affect the comparison of sentiment across news, and it is possible to directly extract the relative relevance of sources.

3.2 Sentiment Analysis

3.2.1 Stanford CoreNLP

Stanford CoreNLP is a Java-based software toolkit developed and maintained by the Natural Language Processing Group at Stanford University. It performs several NLP tasks, such as Part-Of-Speech tagging, Named Entity Recognition, bootstrapped entity learning, sentiment analysis, and Open Information Extraction. At the current stage, the functionality is available in Arabic, Chinese, English, French, German, and Spanish.

The methodology here implemented is based on the Annotator module included in the API. This uses a deep learning model of recursive fashion, trained on the supervised Stanford Sentiment Treebank data set. The key feature of the application is the capability of representing the whole sentence structure and relate the sentiment to various components, rather than using the more classical bag-of-words methodology, that lacks the analysis about words' order. The module and the application is presented in the publication by Socher et al. (2013). They explain well the challenges that sentiment analysis has been facing over the last decades when it comes to classifying multi-categorical data, or short text such as tweets. These barriers derive from the incapability of incorporating the word order in the semantic analysis: the authors hence propose the adoption of the new model called Recursive Neural Tensor Network (RNTN) combining some features of Recursive Neural Network model with others from Matrix-Vector RNN. In fact, the first one allows to keep the number of parameters to estimate relatively low and does not scale it with the size of the vocabulary, while the second provides a good structure to create direct non-linear interaction between input vectors.

Once again, the sentiment analysis is carried out at sentence-level and is then aggregated over documents. Every parsed element is scored on a range 0 to 4, and the within-document variance and polarity are computed.

[6] https://www.ravenpack.com/blog/pairs-trading-news-analytics-deutsche-bank/
[7] https://www.ravenpack.com/research/mean-reversion-pairs-trading-strategy/

3.2.2 Loughran-McDonald Master Dictionary

The second methodology for assessing the sentiment embedded in the news and tweets draws upon the work by Tim Loughran and Bill McDonald from University of Notre-Dame. They presented a combination of publications, Loughran and McDonald (2011) and Loughran and McDonald (2014). They aim at tailoring the text analysis of finance-related documents and assessing companies' official documents readability, to be intended as ease of understanding, and its implication on investors behaviour. The scope of the second academic contribution is partially related to the goal of this project, as they look for new proxies and indicators to extract the ambiguity / readability of Form 10-K, and they relate these measures to the subsequent stock volatility.

The key element for the presented methodology is the Loughran-McDonald Master Dictionary, the core tool used in 2011 paper to examine whether significant statistical correlation exists between the use of negative words and the financial results disclosed by the management. The findings show that the classification of 10k Form filings using the Dictionary lists has a strong correlation with stock price behaviour subsequent the information release. This significance does not persist across time as the information are eventually incorporated in the investors' expectations.

From the technical point of view, the dictionary contains seven lists, identifying different investment attitudes: Negative, Positive, Uncertainty, Litigious, Constraining, Superfluous, Interesting. These are based on *2of12inf dictionary* [8] together with business-specific terminology from EDGAR database [9]. It is interesting to notice that the authors opted for inflections rather than stemmed tokens in order to better match the expressed sentiment with the listed words. The dictionary is updated each year with the new relevant terms at industry-specific and regulatory level.

The Loughran-McDonald approach is replicated here on every news and tweet collected, using only the first two set of word lists. The focus indeed remains on locating negative and positive terms, with the raw count computed separately at sentence level, and then re-normalized by the number of tokens in each string. This provides additional granularity for the analysis as it is possible to measure within-document differences, beside the ones across documents and over time. Thanks to this decision, it is possible to develop features describing not only the average negative and positive sentiment expressed in the document, but also compute proxies that provide an idea about the distribution of sentiment in the content. With this regard, variance and polarity are computed: the former one uses the average sentiment of the sentences as input, while the latter looks at the difference between the most positive sentence and the most negative one.

The nature of this sentiment index implies that the pure sentiment index features assume values between 0 and 1, while other measures such as Polarity are not formally bounded.

[8] http://wordlist.sourceforge.net/12dicts-readme.html

[9] https://www.sec.gov/edgar.shtml

3.2.3 SentiStrength

The third and last sentiment analysis feature is using SentiStrength which is in particular created to analyze short and informal text. Thus, it is perfectly suited for the use of social media texts like tweets. Various applications led to corresponding applications as Thelwall et al. (2010), Thelwall et al. (2012), Thelwall and Buckley (2013) and Thelwall and al. (2013).

SentiStrength was designed for English but can by now handle a variety of different languages. The generated output contains two numbers measuring respectively the negative and positive sentiment in the text. The rationale behind this design choice is that humans also perceive and process positive and negative sentiments or emotions in parallel. Two scores can represent the mixed sentiments of texts more accurately. Scores span from -1 (not negative) to -5 (extremely negative), and from 1 (not positive) to 5 (extremely positive).

Social media text content has its very own characteristics and, besides a general lack of proper spelling or grammar, is likely to contain abbreviations, emoticons (e.g. 'smileys') and truncated sentences. SentiStrength was trained on posts or comments of the social media platform Myspace. It is also dictionary based but the initial word ratings were updated using classified Myspace comments and a machine learning algorithm. Besides that, the dictionaries account for and correct different abbreviations and misspellings of words and are capable of detecting repeated letters or punctuation as aggravated sentiment (e.g. niiiice $>$ nice, the first one indicates higher acceptance or consent than the latter one). SentiStrength does partially also account for the structure of the sentences and recognizes 'booster' or 'negating' words which either aggravate or invert the sentiment of the subsequent words. SentiStrength is also capable of analyzing the sentiment of 'emojis' or 'smileys' into the general score by using a respective dictionary for that purpose.

The sentiment of each sentence, or part of a sentence, is then defined by the most positive and negative sentiment score instead of some kind of aggregation method. Therefore, it is very reasonable to use SentiStrength as an additional sentiment feature because of its specific design for online text.

SentiStrength is provided freely for academic usage but also has a commercial version. It is only available in Java but a Python-wrapper was used for this specific application.

3.3 Sentiment Indices

The extracted sentiment features have been engineered further in order to deal with the nature of the data. The creation of simple and exponential moving average time-series on the average sentiment, the index variance and polarity are helpful for:

- Dealing with the binary structure of the time index : it is only possible to look at the financial returns in hours of market activity, and many documents would be lost if the analysis were to be run only on those published in that specific time frame. The moving averages with longer lags capture the sentiment across the closing hours and the weekends. Usually sensible market information are released off-market hours so this is a reasonable approach.

- Dealing with the noisiness of the data: by its nature, observations come from different sources that tend to interpret and present facts as well as opinions in a specific way, grammatically and semantically. The document collection is also constructed with standard retrieval procedures that allow for less relevant observations to be stored. The use of moving averages, coupled with the weighting of sentiment by importance measures, supports the reduction of the noise.

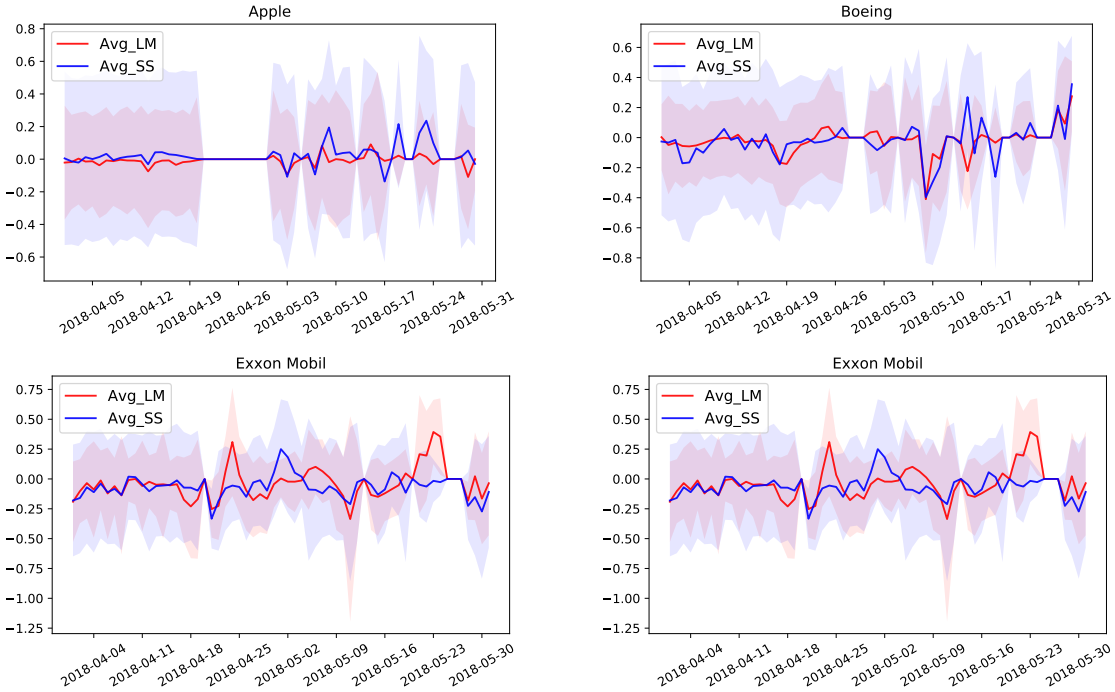


Figure 2: Sentiment of Tweets per entity over time

Figure 3: Sentiment of News per entity over time

4 Modelling And Results

4.1 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) models are a class of statistical models to analyze and forecast time series data. Models are specified by three parameters: (p, d, q), the major components of ARIMA models. These components are:

- Autoregression (AR): This component is referring to the use of past observations in the regression equation of the time series. Its corresponding parameter p specifies the number of lags used in the autoregressive modelling.
- Integrated (I): In many cases the time series might be non-stationary in which case it is common to use the difference of raw observations (i.e. subtracting the current observation from previous values d times) in order to make the time series stationary. Parameter d refers to the degrees of differencing.
- Moving Average (MA): MA parameter q specifies the degree of dependency between current observation and the residual errors from a moving average model applied to lagged observations.

Autoregression, differencing and moving average components in ARIMA model can be described in a linear equation below

$$Y_t = c + \phi_1 y_{d,t-1} + ... + \phi_p y_{d,t-p} + \theta_1 z_{t-1} + ... + \theta_q z_{t-q} + z_t, \quad (1)$$

where Y_t is the current time series observation, $y_{d,t-p}$ is d-differenced value of the time series p steps ago. z_{t-q} is the random noise of the value of time series q steps ago. c is a constant and ϕ's and θ's are the model's fitting coefficients.

Below, ARIMA model is used to fit the minutely close prices of 4 different stocks, where `auto.arima()` in R is used to find the best ARIMA model parameters to minimize the Akaike's Information Criteria (AIC).

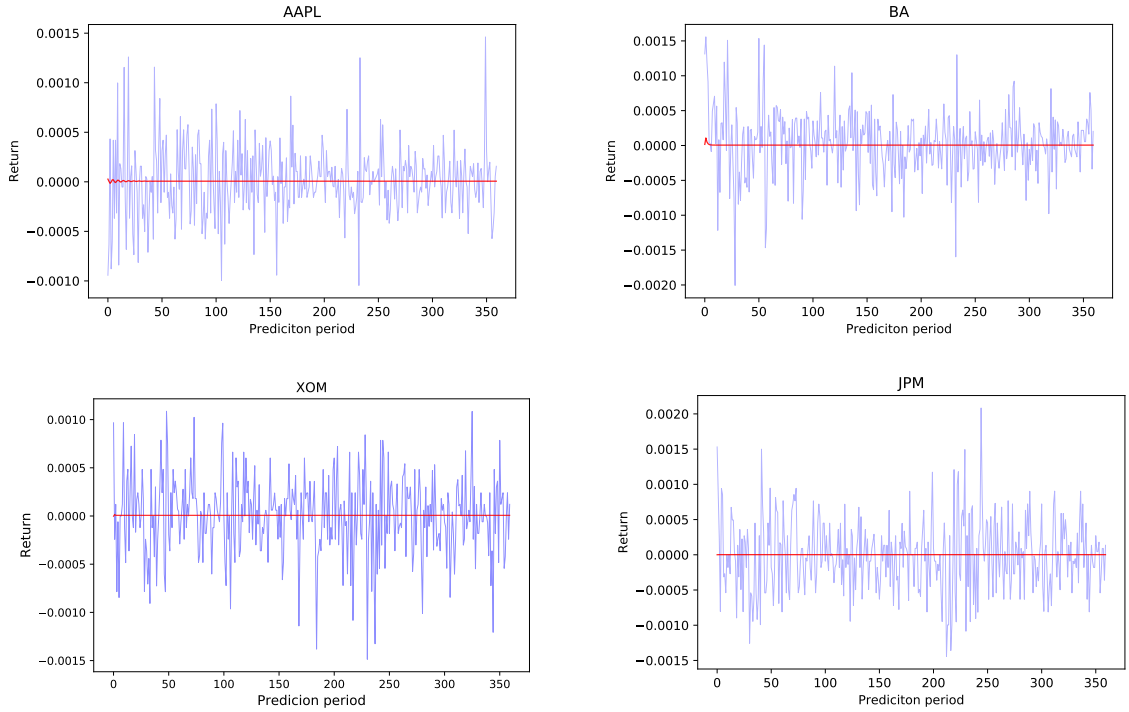


Figure 4: ARIMA prediction versus actual time series

4.2 ARIMAX

As shown in the figure 4, the pure ARIMA model seems to offer a delayed version of the original time series, lacking the ability to predict the pattern. With this regard, one limitation of ARIMA model is that it purely depends on the past values and their historical patterns. Therefore it is not able to explain the underlying time series mechanism based on other covariates. This problem can be solved by using ARIMAX model which besides the historical patterns in the time-series, also takes into account the external covariates. As it can be seen below ARIMAX model have all ARIMA terms plus covariates matrix $\vec{X}_t$.

$$Y_t = \beta \vec{X}_t + c + \phi_1 y_{d,t-1} + ... + \phi_p y_{d,t-p} + \theta_1 z_{t-1} + ... + \theta_q z_{t-q} + z_t, \tag{2}$$

The procedure followed for the ARIMAX implementation is consistent with what will be also carried out for the Machine Learning model in the next section. Firstly, the optimal ARIMAX model parameters (p, d, q) are found through a grid search that compares the different parameters combination and returns the optimal one based on AIC score. In order to deal with severe multicollinearity issues of the covariates, the Variation Inflation Factor ($VIF = \frac{1}{1-R^2}$) method

is applied, to ensure that the final independent variables will have a sufficient degree of linear independence. The selected features are based on the $VIF < 1.5$ criteria. The final outcome of this ex-ante analysis is a final matrix of approximately 30 covariates, depending on the stock.

A 10-fold cross-validation is then implemented with such a dataset to evaluate the forecasting accuracy and avoid over-fitting. The technique depicted below allows to preserve the time series structure of the data-set and also evaluate the performance in a coherent way.

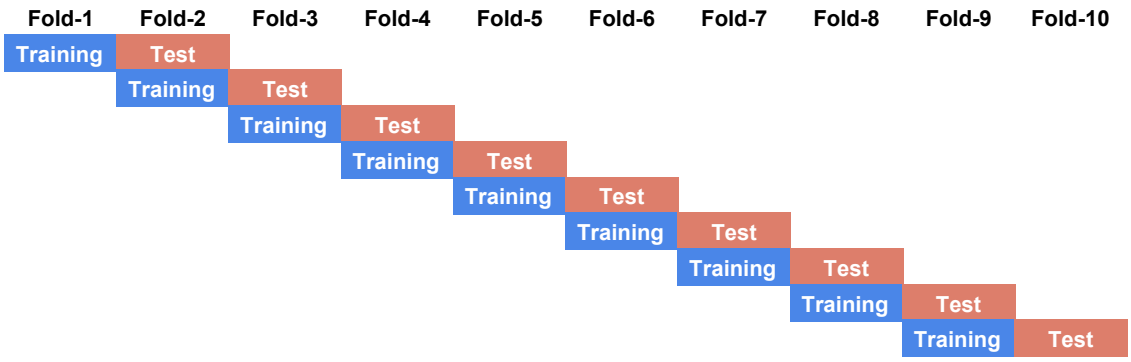


Figure 5: Cross-Validation scheme for time series

In fact, the same amount of training observation is used at each fold and the loss (RMSE) is recorded over five different time windows :

$[0, 5]$, $[6, 10]$, $[11, 30]$, $[31, 60]$, $[121, 360]$ minutes.

ARIMAX - RMSE					
Company	0-5min	5-10min	10-30min	30-60min	2-6h
AAPL	0.0106	0.0093	0.0097	0.0105	2.0525
BA	0.0007	0.0008	0.0009	0.0013	0.0063
XOM	0.0005	0.0006	0.0005	0.0006	0.0010
JPM	0.0016	0.0014	0.0011	0.0011	0.0010

4.3 Machine Learning

We tried to increase the predictive power of the features we constructed by running an Extreme Gradient Boosting algorithm (XGBoost). For each stock, we tuned the model via the previously described cross-validation with fixed train windows. Given the complexity of the model, we decided to run a random search instead of a more complete grid search to reduce the number of combinations of parameters evaluated. In order to reduce the potential overfitting problem of our algorithm we constrained the returns to be between -0.003 and $+0.003$ (which only affected a limited number of observations). Moreover, since the algorithm uses a constant to compare loss values and gives some margins when looking for branch split points, we had to scale the target variable by a factor of 10,000 to make sure that the algorithm was actually learning.

The table below shows the loss associated to every stock for the different time windows. The XGBoost model performs better when trained on the time series of Apple and Exxon Mobil, but it shows a higher root mean squared error when trained on Boeing and JPMorgan Chase.

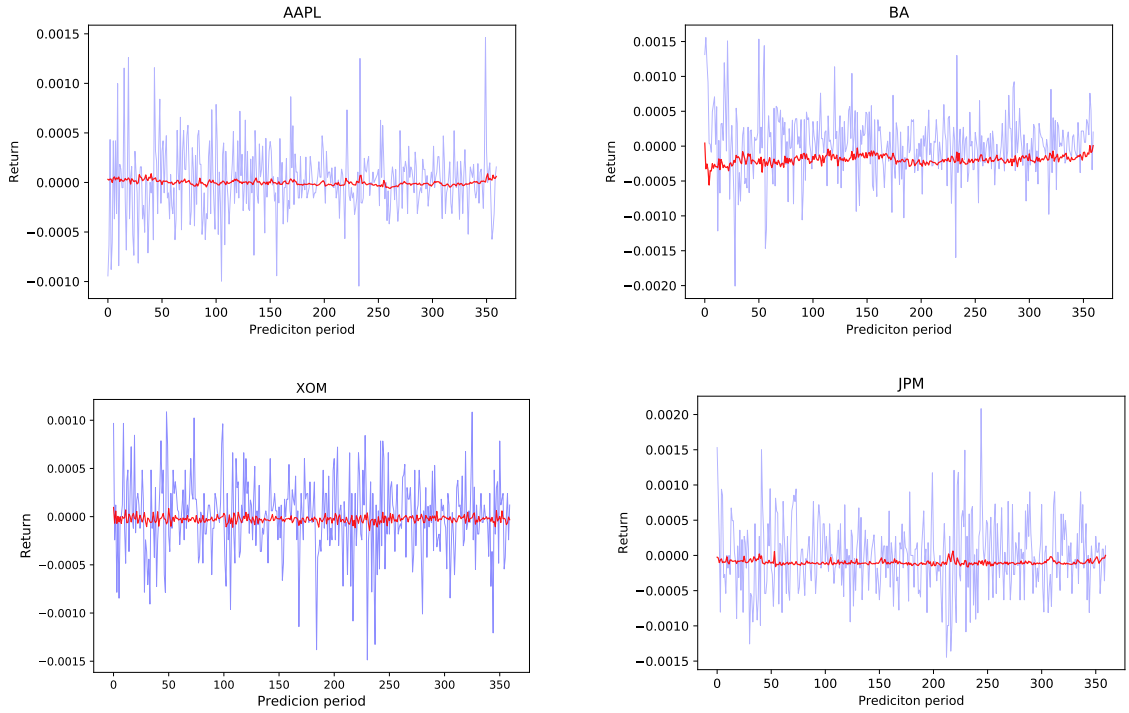


Figure 6: ARIMAX predictions versus actual time series

XGBoost - RMSE Complete Dataset					
Company	0-5min	5-10min	10-30min	30-60min	2-6h
AAPL	0.0015	0.0009	0.0008	0.0007	0.0004
BA	0.0005	0.0008	0.0027	0.0026	0.0348
XOM	0.0004	0.0005	0.0004	0.0003	0.0014
JPM	0.0107	0.0146	0.0125	0.0176	0.0298

In order to take into account for the asymmetry of financial data and available news, we build two different models based on a complete and on a sparse matrix of sentiments. In the former, the interval between consecutive news if filled with the sentiment of the last recorded news (forward filling), in the latter a sentiment appears only when a news is recorded and is zero otherwise.

The following graphs show a comparison between our predictions and the real values in the last 90 percent of our time series (2000 observations).

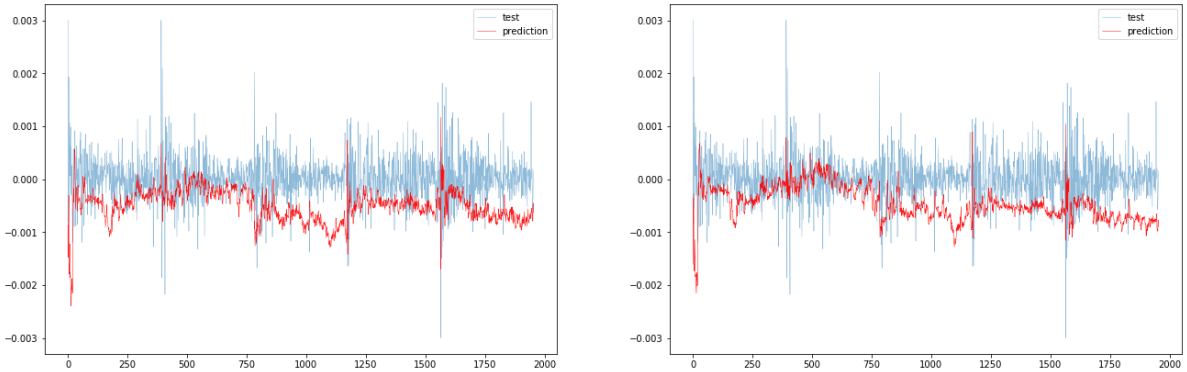


Figure 7: Apple XGBoost test prediction - Complete on left, Sparse on right

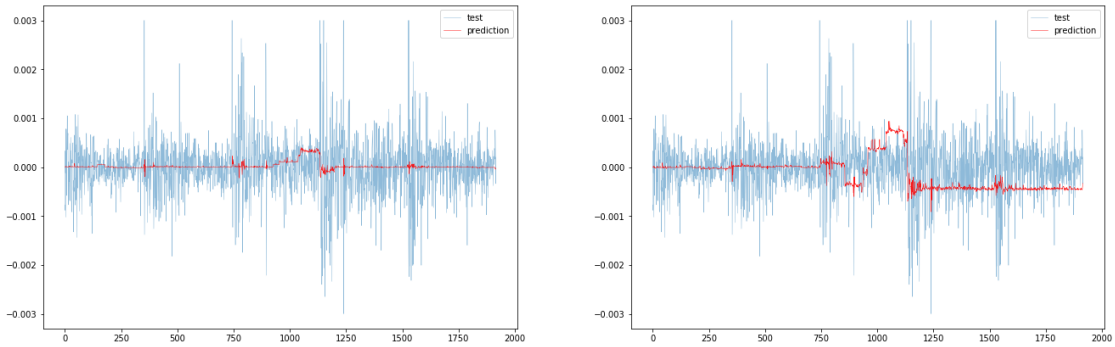


Figure 8: Boeing XGBoost test prediction - Complete on left, Sparse on right

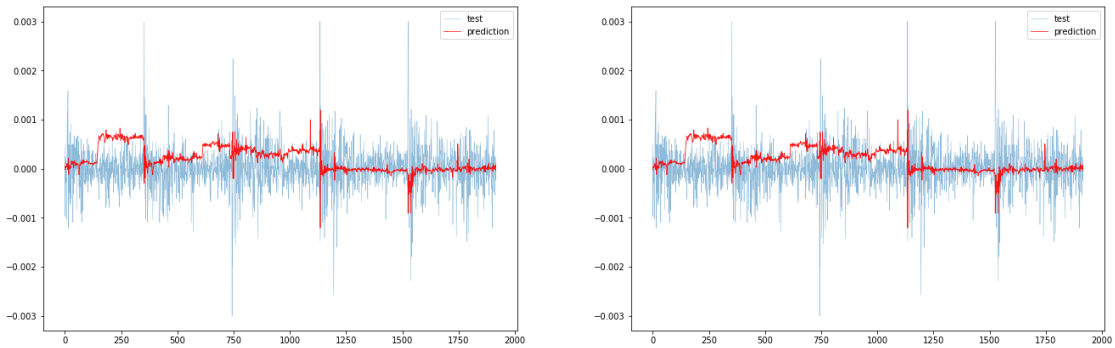


Figure 9: Exxon Mobil XGBoost test prediction - Complete on left, Sparse on right

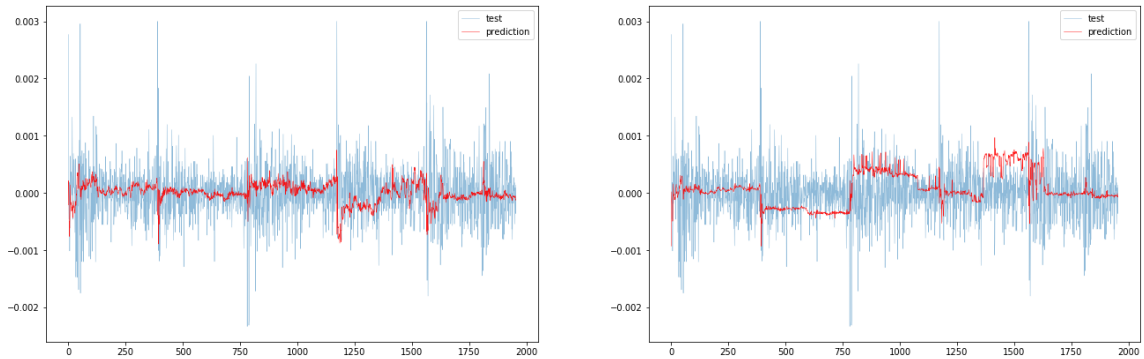


Figure 10: JPMorgan Chase XGBoost test prediction - Complete on left, Sparse on right

We run a features importance check on the tuned models and only lagged values of prices and volumes appear to be the most relevant features in all models. The fact that none of the technical indicators nor the sentiment indicators appear to be particularly relevant is probably due to the white noise behavior of the returns.

5 Discussion

A reasonable consideration about the results presented above is related to the structure of the data-set itself. The minute-level observations allow for high granularity, increased number of data points for the technical indicators. At the same time, news have a more sparse availability than financial data. Therefore, a generally larger time span for the analysis could be beneficial in order to construct a more fine-grained observation structure with the possibility to aggregate sentiment at different time periods (10, 30, 60 minutes). There might be a delay in how markets participants absorb and digest the additional information represented by the sentiment indices. Loughran and McDonald (2014) indeed did not consider company returns for the same day of 10-K form filings. They claimed that it is impossible to divide signal and noise in this situation.

Sentiment could be extracted from a larger variety of news outlets beside those used in this project. More generalist news would help to incorporate general bearish or bullish sentiment which might impact the performance of the general American market. This is specifically referred to as β or market factor, in-line with the Capital Asset Pricing Model presented by Sharpe (1964). Financial markets are relatively sensible to policy changes such as trade agreements, geopolitical tensions, and fiscal and monetary measures. Some of the websites that could be sourced for the above information are CNN, The Guardian, and Reuters. Moreover some of the companies are moved by factors related to other asset classes beyond equity. Fxstreet and Safeheaven [10] are two informative news outlets in this regard.

Furthermore, the possibility to train the sentiment analysis algorithm on supervised data, as in the case of Stanford CoreNLP, would provide a greater accuracy in extracting relevant information and separate signal from noise. Strictly related to news and tweets, techniques of sentiment analysis relying on more sophisticated tools could be applied. Named Entity Recognition algorithms would allow the captured sentiment to be linked to the correct entity of interest, further improving the accuracy. It exists also the possibility to expand sentiment analysis to processed outputs of NLP techniques, for example looking at the sentiment attached to topics modelling outcome for each article.

From the classical econometric point of view, the presence of multi-collinearity is a major issue when trying to use this data setup. The VIF methodology reported above is leading to results that are consistent with the Principal Component Analysis, and allow to safely remove multi-collinear variables. The two approaches could be combined in order to reduce the computational complexity of the algorithm. Furthermore, a variable selection could be implemented within the machine learning approach too, with the aim of reducing computational complexity. A similar procedure could be applied to the machine learning model, to reduce complexity and computing time. Finally, further cross-validation methods could be explored, and the performance evaluated on the back of the work presented above.

[10] fxstreet.com, safehaven.com

6 Conclusion

In summary, a pipeline for news collection, sentiment analysis using three different methods and predictive modelling on four stocks have been created. Five different websites plus Twitter have been used for news collection and their sentiment have been extracted by Stanford-NLP, Loughran-McDonald and SentiStrength sentiment calculator. Each of these methods have their own advantages. Stanford-NLP has higher accuracy because of its deep-learning method of sentiment analysis and the fact that it was pre-trained saving training time. The Loughran-McDonald dictionary has finance-relate terminologies and, therefore, is a more finance-specific sentiment calculator. SentiStrength has the advantage of analysis of short informal texts like tweets and emojis.

Three different predictive models were tuned and compared in terms of forecasting power. ARIMA modelling showed that the lagged-prices and lagged-volatilities contain very little forecasting power consistent with the efficient market theory. ARIMAX models were more capable of picking up actual trends but still far from ideal. The XGBoost machine learning approach showed slight improvement in terms of RMSE of the forecasted signal compared to ARIMAX models. However, even in XGBoost models sentiment-related features did not exhibit relevant intraday forecasting power.

7 Bibliography

References

Bordino, I. and al. (2012). Web search queries can predict stock market volumes. *PLoS ONE*, e40014(7).

Bordino, I. and al. (2014). Stock trade volume prediction with yahoo finance user browsing behavior. *IEEE 30th International Conference on Data Engineering*.

Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2017). Fine-grained sentiment analysis on financial microblogs and news. *International Workshop on Semantic Evaluations*, (11):847–851.

Ferguson, N., Philip, D., Lam, H., and Guo, J. (2015). Media content and stock returns: The predictive power of press. *Multinational Finance Journal*, 19(1):1–31.

Joshi, K., Bharathi, H., and Rao, J. (2016). Stock trend prediction using news sentiment analysis. *CoRR*, abs/1607.01958.

Kerney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, (33):171–185.

Lee, H. and al. (2014). On the importance of text analysis for stock price prediction. *LREC*.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *CIKM 09, Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13:1–26.

Nassirtoussi, A. K. and al. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, (41):7653âĂŞ7670.

Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance*, 19:425–442.

Sinha, N. R. (2010). Underreaction to news in the us stock market. *Quarterly Journal of Finance*, (6).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Conference on Empirical Methods in Natural Language Processing EMNLP*.

Sul, H. K., Dennis, A., and Yuan, L. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3).

Thelwall, M. and al. (2013). Damping sentiment analysis in online communication: Discussions, monologs and dialogs. *CICLing*, II(7817):1–12.

Thelwall, M. and Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, (64):1608âĂŞ1617.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, (63):163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, (61).