Politecnico di Torino

Corso di Laurea in Ingegneria Matematica

A.A. 2021/2022

Sessione di Laurea Dicembre 2022

Covid-19 and the Financial Markets: Analysis of the Correlation Between Tweets Sentiment in the United States and the Return of a Portfolio during the Market Crash

Relatori:

Prof. ssa Tania Cerquitelli (Politecnico di Torino)

Prof. ssa Alberta Di Giuli (ESCP Business School)

Candidato:

Riccardo Vellano

To i nonni.

They have taught me respect, humility and hard work. They ignited and nurtured my passion and astonishment for nature and mathematics. Sharing love, they never missed the opportunity to teach me about the past they had lived, giving me strong roots to build my life on. To our memories together and to the future moments. As my grandfather used to say "Per aspera, ad astra".

To mamma, papà and Gaia.

Always supporting me when I am pushing beyond my boundaries and guiding me through the most difficult decisions. Nothing of the accomplishments I am most proud of would have been possible without them.

To I Fantastici.

To our long-lasting friendship started 11 years ago. To our amazing adventures together discovering the World, knowing that we will always be there for each other.

To all my friends and family.

To the friends of a lifetime and to the recent friendships. I have learnt that it is never too early to board a plane for an adventure with a yellow coated friend you have just met. I feel extremely lucky to have you all by my side, always being close to me during these years. A special thank you goes to everyone who supported me during these challenging months of working and writing. Especially, thanks to my cousin and Parisian flat mate Francesco and to my other 90%, Silvia.

"Chi ha tempo, non perda tempo"

Table of Contents

1. Abstract

This Master Thesis research investigates the possibility of predicting markets fluctuations during the first months of the COVID-19 pandemic in order to enhance the returns of a portfolio constituted of sectorial ETFs. As previously researched by many authors and argued by Gu and Kurov (2020), investors rely on social media sentiment and information to make investment decisions. Over the years, sentiment analysis of social media posts for stock market prediction has attracted a lot of interest and this was demonstrated by many papers researching the topic. As argued and demonstrated by Bollen et al. (2011) in "*Twitter mood predicts the stock market"*, a pivotal paper in this field of research, it is possible to predict stock market fluctuations by analysing the sentiment on Twitter. To this purpose, in this research around 1.4 million tweets are downloaded from the social network via an Academic Research Developer account. All tweets were posted between 23rd February 2020 and 31st May 2020 from the United States, in English and they focus and comment on the topic of COVID-19. An exploratory analysis of the dataset shows that the tweets downloaded express feelings, opinions, comments or report news related to the pandemic, but they were chosen to be posted by generic user with the specific goal of having an unbiased dataset. Via the use of a cutting-hedge Natural Language Processing algorithm architecture called Transformers, the sentiment is extracted for every tweet. It can be positive, negative or neutral based on the words used in the tweets and the feelings expressed. In parallel, the research focuses on sectorial Exchange Traded Funds, or ETFs, which are funds traded on the markets that aim at replicating the performance of certain securities. In this case, the ETFs are used to replicate the performance of specific sectors which were particularly influenced by the pandemic: the industrial, financial, healthcare and technological sectors. The returns of these ETFs represent the returns achieved by an investor willing to have a financial exposure to the aforementioned sectors. The 30-minutes intervals performance of the ETFs is linked to the sentiment expressed on Twitter during the same timelapse to create a dataset to train algorithmic models for prediction. The study of precedent research inspired the use of supervised learning models such as k-Nearest Neighbours and Support Vector Machines to forecast directional fluctuations in the ETFs price. It is shown that the kNN algorithm performs best with around 58 % accuracy on the training dataset. By using the directional forecasts performed by the model on test dataset composted of May 2020 tweets and ETFs performance, it is shown that these predictions help improve the returns of a portfolio invested in the Industrial Select Sector SPDR Fund (XLI) ETF during the COVID-19 from 7.33 % to 15.84 %. Therefore, an improvement of 8.51 % in the portfolio return is achieved when using insights produced by the sentiment analysis of tweets combined with supervised machine learning algorithms. All the algorithms are coded in Python and have been developed independently with the use, where indicated, of pre-compiled Python libraries especially for the Transformer architecture, SVM and kNN models. In conclusion, this Master Thesis research suggests the possibility of sentiment analysis-based yield

enhancement strategies taking advantage of the information shared on Twitter by its users, rooted on the so called Wisdom of Crowds, Surowiecki (2004).

2. Literature Review

2.1 COVID-19 Pandemic and its economic impact

The outburst of the Covid-19 pandemic represented an unprecedented challenge for the world's population. The disease that originated in the Chinese city of Wuhan in December 2019 quickly spread abroad reaching all countries and causing a pandemic. Governments were forced to impose strict restrictions and shut down all activities considered as not fundamental. Moreover, a series of quarantines and lock-down measures hit many countries: among these Italy went on full lockdown on March 10th, 2020 shortly followed by most European countries and the US on March 19th, 2020. Since many businesses could not sustain paying employees without having inflows due to lockdown measures, many people found themselves unemployed – in the US the unemployment rate reached 20%[1]. This led to a further decrease in consumption and in a dramatic decrease of expected future cash flows for companies, hence a significant decrease in enterprise value with the subsequent correction on the markets.

Mazur et al.[2] (2020) compare the Dow Jones Industrial Average (DJIA) losses during that period to the well-known market crashes of Black Tuesday on the 28-29th October 1929, and of Black Monday on the 19th October 1987 when the DJIA index lost respectively 24.5% and 22.6%. Indeed, during the four days March 9th, 12th, 16th and 23rd 2020 the same index lost overall roughly 26%. The heavy losses, however, were not homogeneous on the companies. Some sectors suffered heavier losses than others, while some other benefitted from the pandemic. In particular, Mazur et al. find that the crude oil sector was hit the hardest and companies lost over 60% in a day, also experiencing extremely high volatility. Similar losses were suffered by the real estate, hospitality and entertainment industry. On the other hand, all those companies providing services and goods that could be consumed at home during lockdown periods saw significant increases in their market capitalization. That is the case of companies in the healthcare, food, software, technology and natural gas sector.

Governments and institutions tried to quickly respond to the spread of the virus with measures such as travel bans, lockdowns and stimulus packages in order to offer support and smoothen the impact on the affected population. Central banks also reacted by reducing interest rates so that governments could borrow money at advantageous rates, hence helping the economy and preventing a collapse. As argued by Narayan et al. [3] (2020) overall these measures worked to cushion the effect of COVID-19 on the stock market.

[1] https://faculty.fuqua.duke.edu/~charvey/Audio/COVID/COVID-Harvey.html
[2] Mazur, M., Dang, M. and Vega, M. (2020). COVID-19 and the March 2020 Stock Market Crash. Evidence from S&P1500. *Finance Research Letters*, 38(101690), p.101690. https://doi:10.1016/j.frl.2020.101690
[3] Narayan, P.K., Phan, D.H.B. and Liu, G. (2020). COVID-19 lockdowns, stimulus packages, travel bans, and stock returns. *Finance Research Letters*, 38, p.101732. https://doi:10.1016/j.frl.2020.101732

Several econometric researches have been conducted to analyse different aspects of the markets and their reactions to such a shock. In *Resiliency of Environmental and Social Stocks*[4], Albuquerque et al. (2020) explore the possibility of lower losses for highly environmentally and socially (ES) compliant stocks in difficult market conditions. The paper highlights the extraordinary character of the period analysed, which saw an unprecedented spike in the number of commercial and industrial loans in banks' balance sheets. As already specified, Central Banks allowed for a further decrease in interest rates, which were already at historical minima, therefore the number of bond issuance increased, especially for high yield rated bonds.

The authors, however, make an important distinction between the COVID-19 crisis and the Great Recession. The first and most striking difference is the nature of the cause of the two crisis: on one side the Great Recession was driven by internal economic causes such as the subprime crisis, which fundamentally led to a mistrust in the financial system. On the other hand, the COVID-19 crisis can be defined as an "exogenous and unparalleled stock market crash" because the cause was external to the financial system. The lack of possibility to forecast and prepare for such a crisis caused a very strong correction in the market which lasted for an extremely limited amount of time. On the contrary, the Great Recession lasted for 2 years and the severe drops in the markets were more spread during time. This distinction, also made by Mazur et al. (2020), underlines the fact that the COVID-19 crisis represents a great opportunity to study theories in a less noisy and more concentrated market environment, where all the major causes of stock crashes were linked to the pandemic. Indeed, Albuquerque et al. (2020) conclude that "the first quarter of 2020 was an extraordinary time for U.S. stock markets: first a calm period before the storm, then the fastest collapse ever, followed with a vigorous rally, all related to the unfolding of an unexpected, exogenous, health pandemic".

For the purpose of this research, the aforementioned paper constitutes an extremely interesting source in the choice of the timeframe this research intends to analyse. COVID-19 was already present in China at the end of 2019, however markets had not shown the minimum sign of fear for the virus. Only after the virus expanded internationally and was first discovered in North Italy, the Western cultures started being worried. However, it was after the first sectorial lockdown in some municipalities in Lombardy and Veneto that investors began realizing what could happen to the whole World once the virus had expanded. Therefore, the paper considers as starting date February 24th, 2020, which is the first trading day after the announcement[5] made by the former Italian Prime Minister Giuseppe Conte about the lockdown of those municipalities. It is to be noted that the S&P500 peaked on February 19th, and due to the previous considerations in this research the date of February 24th, 2020 will be identified as the start date.

[4] Albuquerque, R., Koskinen, Y., Yang, S. and Zhang, C. (2020). Resiliency of Environmental and Social Stocks: An Analysis of the Exogenous COVID-19 Market Crash. *The Review of Corporate Finance Studies*, [online] 9(3), pp.593–621. https://doi:10.1093/rcfs/cfaa011

[5] The announcement was made on February 23rd, 2020. These dates also coincide with the "fever" period identified in Ramelli and Wagner (2020)

As mentioned above, most of the economic shock was caused by the impossibility to sell or produce goods because the related business activities conflicted with the regulations put in place to limit the spread of the virus. Fahlenbrach et al. (2020)[6] argue that the crisis should affect less those companies which have a more flexible cost structure, hence more variable costs, which can be reduced in case of a reduction in productivity. This flexibility can also be linked to the debt structure. Their finding suggests that more flexible firms achieve a 26% lower stock price drop compared to the same industry peers.

Overall, the analysis of these papers show the extraordinary market conditions that the pandemic provoked. Therefore, it can be argued that the COVID-19 crisis constitutes an event with very little noise compared to previous crises and in this environment it is possible to study the correlation between sentiment expressed online through social media and the markets.

2.2 Population Sentiment during COVID-19 and Twitter sentiment analysis for financial purposes

From a psychological perspective, the pandemic heavily affected individuals. The prevailing feelings were fear, high vulnerability and insecurity because the population was dealing with a deadly illness which was mostly unknown and for which there was no certain cure nor vaccine. People found themselves suddenly confined at home with rare opportunities to exit except for essential reasons. Many people suffered from anxiety and depression due to the many losses and hospitalizations in ICUs of their closest ones, without, most of the times, having the possibility to visit them.

These conditions made people spend more time at home and on social media, where their thoughts would be shared incorporating the users' feelings as well. The intention of this research is to use sentiment expressed on Twitter to investigate its correlation with the markets' performance and the possibility to create a predictive model to improve an equity portfolio's performance.

The literature on this topic is already rich but it still remains a field of research and cutting-edge businesses invest in this – the company SESAMm[7] is an example. It uses artificial intelligence to gather financial insights from news articles and sells them to global investment firms. As reported by Baker et al.[8], COVID-19 represented a shock for the markets which was unexperienced even during previous and more lethal pandemics in the 20th century. In "*The Unprecedented Stock Market Impact of COVID-19*" the authors analyse

[6] Rüdiger Fahlenbrach, Kevin Rageth, René M Stulz, How Valuable Is Financial Flexibility when Revenue Stops? Evidence from the COVID-19 Crisis, *The Review of Financial Studies*, Volume 34, Issue 11, November 2021, Pages 5474–5521, https://doi.org/10.1093/rfs/hhaa134

[7] https://www.sesamm.com/

[8] Baker, S.R., Bloom, N., Davis, S.J., Kost, K.J., Sammon, M.C. and Viratyosin, T. (2020). *The Unprecedented Stock Market Impact of COVID-19*. [online] National Bureau of Economic Research. Available at: https://www.nber.org/papers/w26945

the first article in the newspapers after a day of stock market losses higher than 2.5%: they categorize the article based on its content. By using this criterion they analyse more than 1100 stock market jumps in a period from January 1900 until April 2020, these represent only 3.5% of all trading days in the period, but overall 47% of the total squared daily return variation in the same period. During these 120 years, there were different pandemics such as the Spanish Flu in 1918-1919, the influenza pandemics in 1957-58 and 1968 and the latter COVID-19 pandemic in 2020.

Authors do not find strong daily market swings related to the previous infectious diseases: for instance there were 23 daily stock market jumps from March 1918 to June 1920, of which none relates closely to next-day articles about the Spanish Flu. Similarly, there were 9 jumps in 1957-58 and one in 1968, none of them related to articles on the pandemics development on the Wall Street Journal. On the contrary, since 24th February 2020 until 24th March 2020 out of 22 trading days, there were 18 market jumps, symptom of high volatility and stressed market conditions. Authors affirm that if the observation period is extended to the end of April 2020, there were 27 jumps, of which 23 were attributed by the New York Times and the Wall Street Journal either to the coronavirus pandemic development or to governments responses. It is important to notice that the swings happened in both directions.

According to the researchers, various factors can contribute to explain this behaviour. The first argument is that starting from February the pandemic's developments started to strongly dominate the newspapers coverage and discussions also in the economic sector. High volatility caused a surge in uncertainty which is clearly represented in the volatility during that period. This was caused by the nature of COVID, which was easily transmissible and with a non-negligible mortality rate among those who contracted it. However, the mortality rate was limited to 1/25th of the Spanish Flu at the time of the article[9], hence the answer to such high reaction rate must rely also on other factors.

The Spanish Flu happened in times where the majority (61%) of the employment was in the Agriculture and Manufacturing sector. On the other hand, COVID-19 unfolded in a context where the Service sector is prevailing. This sector requires high contact between individuals who would travel with high frequency, therefore all the governmental policies put in place strongly affected the economical equilibrium. Moreover, news spread with a higher rate compared to the beginning of the 20th century, hence the impact of COVID-19 on the markets results more concentrated and more likely to trigger high stock market jumps. Dependability on cross-border flows of goods and supply chains which often rely on production of different components in various countries has caused a severe disruption in industries. All these reasons, together

[9] Barro, Ursua and Weng (2020) report U.S. excess mortality rate to 0.52 percent of the population from 1918 to 1920, compared to 0.02 percent during the first months of COVID-19 pandemic.

with social distancing policies, whether imposed or voluntary, constitute possible explanations for such an extreme effect of the latter pandemic on stock markets, compared to all the previous pandemics.

The previous document introduced the correlation between newspaper articles and stock market movements. In general, textual sentiment analysis has represented lately a field of research for its possible correlation with the stock market and the possibility to use social media mood to predict the stock market performance has attracted increasing attention from researchers.

Since its foundation in 2006, Twitter has developed to become one of the most influential social media on the Planet. Many investors rely on insights shared on the social media to make investment decisions, as argue Gu and Kurov, 2020[10]. Indeed, they mention the example of when the American Diabetes Association announced worse than expected results for the clinical trial of a drug produced by Novo Nordisk. Despite the association having requested the maximum confidentiality over the shared information until it was officially made public, the attendees soon shared impressions on Twitter causing Novo's price to drop significantly. Many studies involve the use of Twitter for financial purposes, thus evaluating its importance in making investment decisions. Despite the high number of findings achieved so far, this remains an active field of research given the difficulty in assessing the amount of insightful information and the amount of noise carried by the extensive quantity of data that must be analysed.

An example of these studies is the paper "*Corporate Twitter use and cost of equity capital*", in which Al Guindy[11] investigates if posting tweets reduces the cost of capital. In particular, the author argues that companies tweeting information about their financial results and core business achieve lower costs of capital by reducing the asymmetry of information. The author investigates two hypotheses: the first states that firms adopting Twitter have a lower cost of equity than those not using it; based on the author's second hypothesis, among these firms using Twitter the ones that benefit the most should be the small firms. These are normally subject to higher asymmetry of information since they have a lower analyst coverage compared to bigger and well-known companies and, therefore, investors request a higher return for a higher risk given by increased information asymmetry. Analysts incur in a "cost" for each firm they follow (Harford et al., 2019[12]) given by the limited amount of time and attention they possess. Companies posting on Twitter may help reduce the difficulty in finding information for analysts and may also attract attention, reducing consequently the "cost" experienced by analysts. Thus, this would mean a higher analyst coverage and more easily available information on Twitter, therefore the cost of equity capital would decrease.

[10] Gu, A. professor of finance C. and Kurov, P. of finance A. (2020). Informational Role of Social Media: Evidence from Twitter Sentiment. Journal of Banking & Finance, p.105969. https://doi:10.1016/j.jbankfin.2020.105969

[11] Al Guindy, M. (2021). Corporate Twitter use and cost of equity capital. Journal of Corporate Finance, p.101926. https://doi:10.1016/j.jcorpfin.2021.101926

[12] Harford, J., Jiang, F., Xie, F., 2019. Analyst career concerns, effort allocation, and firms' information environment. Rev. Financ. Stud. 32 (6), 2179–2224.

In April 2013 the SEC allowed public companies to distribute important news, like quarterly earnings, to the public via Twitter. Somebody could debate that, this information being available on the company website, Twitter should play no major role. However, Twitter's role is the dissemination of information, spread to various users via an algorithm that identifies the main interests of a user and shows related information about companies that the user may not know. The fact that this information is also available on the company's website, implicitly assumes that the public already knows the company and is looking for this information on the internet, so it does not help disseminating knowledge. By testing these two hypotheses via the tweets analysis, the author finds that an effective reduction in equity cost of capital between approximately 20 and 30 bps is achieved by firms posting generic and financial information on Twitter.

Twitter data is used in other applications, such as in the prediction of oil futures volatility in the article written by Lang et al, 2022[13]. However, one of the most relevant is the aforementioned article "*Informational role of social media: Evidence from Twitter sentiment*" by Gu and Kurov[7]. The authors investigate whether Twitter sentiment predicts stocks returns. The main types of previous research around this topic identified by them are of two kinds: the existing studies that focus on a specific type of events evaluating the ability of tweets in forecasting stocks movements related to a specific event and those studies that consider tweets globally and measure the general forecast ability. In their study, the authors use Twitter sentiment about stocks provided by Bloomberg to demonstrate that various companies information can be predicted by analysing Twitter sentiment. The Bloomberg tool providing Twitter sentiment is based on proprietary algorithms that rely on supervised machine learning. A pool of tweets is labelled at first by Bloomberg analysts and then fed into machine learning algorithms recognizing patterns of "bullish" and "bearish" tweets. This trained algorithm is then used to label subsequent tweets and provide the daily sentiment every 24h on the Bloomberg terminal. Therefore, the authors are able to use the daily sentiment of the previous day to investigate if it has a predicting potential for stocks. Two hypotheses are made: tweets have a short-term impact on stocks because they are made by misinformed investors and do not represent the market consensus on the stock, or the tweets contain some financially insightful information which is not incorporated yet in the market price. If that is the case, the movement predicted by the Twitter sentiment will have a long-lasting impact. Through regressions, the authors find that Twitter sentiment predicts analysts' recommendation upgrades, as well as target price increases and higher earnings. It also predicts announcement days returns, as well as IPO under-pricing. These findings suggest that tweets include pieces of information that have not yet been priced in the market and the absence of subsequent reversals in the stock price indicates that this information is mostly reliable. The authors also test their theory by building a

[13] Lang, Q., Lu, X., Ma, F. and Huang, D. (2022). Oil futures volatility predictability: Evidence based on Twitter-based uncertainty. *Finance Research Letters*, [online] 47, p.102536. https://doi:10.1016/j.frl.2021.102536

long-only and a short-only portfolio which they manage using their findings to show that it is possible to obtain significant portfolio returns by predicting the stock movements with this strategy.

These papers and many more about sentiment analysis on Twitter, which will be investigated in the subsequent section, clearly show the existence of noise-hidden knowledge in the tweets. Experienced traders could investigate which accounts provide the most reliable information and focus on those for their investing strategy, allowing for a partial noise reduction. Every user on Twitter contributes with partial information to a higher level of knowledge. This phenomenon can be considered as an example of "wisdom of crowds" (Surowiecki, 2004[14]).

Literature on this subject is particularly rich and the article "*Twitter mood predicts the stock market*" by Bollen et al. (2011)[15] is a pillar, which many others make reference to. In the article, the authors investigate the correlation between emotions expressed on the social network Twitter and the markets. Behavioural economics supports the thesis that emotions can heavily affect behaviour and decision-making processes, therefore the authors investigate the possibility that societies at large may be influenced by general feelings patterns which may lead to market fluctuations.

The authors argue that the Efficient Market Hypothesis presents two main issues: first of all, many studies show that the markets tend not to follow a random-walk pattern, but that it may be predictable. Moreover, while news may not be predictable, some early indicators can be extracted from social media. An example of the use of Twitter to understand the overall sentiment of a Country is represented by "Pulse of Nation", which is a sentiment tracking indicator[16]. The authors investigate "whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time". By performing some dataset cleaning before proceeding with the tweets analysis, the researchers keep only those tweets including words expressing feelings. Thereafter, they employ the software OpinionFinder to measure the level of subjectivity of a tweet and of positivity or negativity it expresses. However, they find that this is not sufficient, hence they create a list of six possible emotional states in which they categorize the tweets: calm, alert, sure, vital, kind and happy. They evaluate the ability of a neural network to predict the DJIA's performance and employ a SOFNN (Self-organizing fuzzy Neural Network), consisting of a five-layers hybrid neural network which self-organizes its neurons' weights in the learning process. The paper concludes that the emotional state *calm* helps predict with an 86.7% accuracy the directional performance of the markets, put in other words it predicts whether the DJIA will increase or decrease with a significative accuracy.

[14] Surowiecki, J., 2004. The Wisdom of Crowds. Random House, New York
[15] Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp.1–8. https://doi:10.1016/j.jocs.2010.12.007
[16] https://www.ccs.neu.edu/home/amislove/twittermood/

Other papers get inspiration from the research conducted by Bollen et al. (2011), among these particularly Mittal and Goel[17] obtain about 75.6% prediction accuracy, highlighting the goodness of Bollen et al.'s research.

The paper "*Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media*" by Chen et al. (2014)[18] represents an extremely valuable source for this research. Indeed, the authors investigate to which extent the opinions expressed on social media such as the website Seeking Alpha mirror the effective markets performance of stocks. The latter's goal is to provide "opinion and analysis rather than news, and is primarily written by investors who describe their personal approach to stock picking and portfolio management" (Seeking Alpha 2012).

According to the paper, 25 % of adults in 2008 in the U.S. relied on investment advice transmitted via social media. This percentage, given the development of social media nowadays, can be considered even more important. As an example, it is sufficient to think about blogs and social media that were born exactly with the intent to discuss financial topics such as Reddit. The tremendous impact these internet communities have on the financial system is clear when analysing some case studies like the ones of BlackBerry or GameStop[19]. This implicates that the paper's assumptions are still valid, if not strengthened by the current situation. On SA (Seeking Alpha) users can either provide opinions on investments by publishing articles, or by commenting other articles.

SA is compared by the authors to a bottom-up knowledge generator such as Wikipedia, for example. The reason for this relies on the way knowledge is created, evaluated and diffused. Indeed, the whole network contributes to the generation of knowledge by writing articles and then commenting them; at the same time, users evaluate articles based on the quality of information contained in it and also on the historical accuracy of the evaluations made by the authors. This is a more democratic knowledge participation and sharing compared to the knowledge produced by investment professionals. Authors argue that it is plausible that large crowds possess insights on companies that may not be available to single company analysts because each individual in the crowd contributes with small pieces of information, sometimes even bringing into the social media some news or comment that was not known by the professionals. Although this might be very rare, however wisdom of crowds can also be preferred to single analysts' opinions because in such big contexts conflicts of interest tend to disappear and nihilate one another, whereas financial professionals have high conflicts of interest. On the other hand, however, it is to be taken into consideration that

[17] Mittal, A. and Goel, A. (n.d.). *Stock Prediction Using Twitter Sentiment Analysis*. [online] Available at: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[18] Chen, H., De, P., Hu, Y. (Jeffrey) and Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, [online] 27(5), pp.1367–1403. https://doi:10.1093/rfs/hhu001

[19] https://www.ft.com/content/f8937ac7-da4a-41a6-bbbf-7b573e8ed72b

professional analysts have higher degrees of visibility in the field than authors on SA. This causes professionals' opinions to diffuse at a faster pace and get incorporated into markets prices quicker. According to Chen et al., should social media importance increase in the financial field, then it would be possible to see opinions expressed on SA, for instance, be incorporated into the markets at similar paces.

The dataset used for their research included all article published between 2005 and 2012 and their comments on the two subsequent days. Overall, these accounted for 97'070 single-ticker articles, plus all the comments. The number of comments to single-stock articles were also used to improve the accuracy of the model. The latter consists of a multiple variable regression based on two independent variables: the fraction of negative words in the article and the average fraction of negative words published in the comments to the article. Indeed, the key finding of this paper is that the predictability of the stock performance increases if the fraction of negative words contained in the article and in the comments is taken into account. The possibility to use these articles and their related comments to predict the stock performance is explained in two ways: the first states that the view expressed in the SA article includes for real some piece of relevant information that has not yet been incorporated in the stock's market price. The second suggests that naïve investors may be persuaded by these articles, hence they would act on the market as suggested by the articles' authors thus provoking the predicted price movement. In general, the authors conclude that they cannot evaluate with certainty which of the two explanations is the correct one and that, probably, the market performance is the result of both factors.

What is also interesting about this paper is the application of the authors' findings to a business example. Indeed, they implement a portfolio management strategy based on the fraction of negative words found in the SA articles related to the stock. Every day, they go long on the bottom 20% of stocks with the least negative words ratio and they go short on the top 20% stocks with the most negative words ratio in the articles and comments. Then, they keep the stock in the portfolio for three months. The strategy leads to around +40 % portfolio return in over 6 years.

Sentiment analysis of tweets has already been implemented in some well-known financial services such as Thomson Reuters Eikon[20] and Bloomberg[21]. In "*The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*" Oliveira, Cortez and Areal, 2017[22] use lexicon-based unsupervised algorithms (such as in Bollen, Mao and Zeng, 2011[10]) to determine the sentiment of tweets and then predict returns of the S&P500 index, as well as some small

[20] https://www.thomsonreuters.com/en/press-releases/2014/thomson-reuters-adds-unique-twitter-and-news-sentiment-analysis-to-thomson-reuters-eikon.html
[21] https://www.bloomberg.com/company/press/trending-on-twitter-social-sentiment-analytics/
[22] Oliveira, N., Cortez, P. and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, pp.125–144. https://doi:10.1016/j.eswa.2016.12.036

market capitalization companies and some industries. A lexicon-based algorithm is an algorithm that classifies texts based on the score that is attributed to every single word in the text. It relies on a precompiled dictionary where positive and negative words are given different scores based on their positive (or negative) meaning. The algorithm assigns scores to the words in the text based on the aforementioned dictionary, then sums all the scores and gives a general sentiment to the tweet based on the final score. A supervised method defining whether tweets are positive or negative is also adoptable but requires manual labelling of hundred of thousands of tweets, before being able to run the algorithm, which makes the implementation of the latter strategy particularly difficult. While using lexicon-based unsupervised models, the authors also take into consideration that bigrams (two consecutive words) better represent the meaning in a text, compared to lexicons based on single words. For instance, in the bigram "debt free" the word "debt" has a positive overall meaning, which would not be understood, if only considering single words.

Having defined the sentiment of the tweets, the authors run different models to characterize the correlation between the stocks' returns and the tweets. These models all represent regression models and include a Multiple Regression, Neural Network, SVM (Support Vector Machine), Random Forest and Ensemble Averaging. The Multiple Regression model assumes a linear relationship between the independent variables and the dependent variable, which in this case does not reveal to be appropriate. This is the reason why it performs poorly, despite having many favourable aspects such as the easy interpretation, the quick learning rate and easy algorithms to implement it. Of all the algorithms, the authors find that the best performing is the Support Vector Machine.

Results show that microblogging sentiment indicators were informative for predicting the S&P500 Index, as well as some industries such as High Technology, Energy and Telecommunication. Also, they revealed to be important predictors for low market capitalization companies. The explanation given for the latter finding, in particular, is that these stocks are most of the times held by retail investors that are quite sensitive to news and general sentiment, hence further from rational investors who hold the majority of big companies. The methodology explained in this paper is extremely helpful in defining the data analysis path followed for this research and possible expected results. An extremely positive aspect of sentiment analysis from social media such as Twitter is that data is easily collected and present in enormous quantity, which is a key aspect for this kind of algorithms. Moreover, it is a strategy that allows cheap and great frequency investment insights which can be customized for specific stocks. Therefore, the proposed methodology, as concluded by the authors, "could be adopted by financial expert systems to support investors in their decisions by providing instant access to social media analytics, such as customized sentiment indicators or predictors".

2.3 Transformers

Briefly summarizing the previous sections of this Literature Review, it has been demonstrated that the COVID-19 pandemic period represents a unique opportunity to observe and verify theories in the market. This is mainly due to the short period in which heavy measures were taken to contrast the virus diffusion, leading to a market crash which has seen very few prior events of such importance and never caused by a pandemic. In an era when social media play an essential role in communication, analysing the correlation between the stock market and social media has risen great interest among the research community. In particular, the objective of this research is to investigate the correlation between the sentiment expressed on tweets and the stock market during the first months of the pandemic outbreak in order to predict the markets' performance. So, the intraday Twitter sentiment will be needed as well as the stock market returns. To obtain the intraday sentiment, a sentiment analysis of the tweets will be performed and in this section the literature behind the methods applied will be analysed.

Previous approaches to this research have seen the use of lexicon-based algorithms to define the sentiment of tweets. However, this approach presented two main issues: the first being the enormous amount of computational power required to pre-process the text using the TF-IDF method and filtering stopwords – these topics will be discussed subsequently in the "Data Analysis" section. Secondly, as pointed out by Oliveira, Cortez and Areal (2017), most of the times lexicon-based algorithms are not capable of extracting the exact meaning of a sentence because they focus on single words, rather than considering bigrams or the whole sentence. These considerations led to the adoption of state-of-the-art algorithms called Transformers.

Transformers were first introduced by Vaswani et al. (2017)[23] in their notorious paper "*Attention Is All You Need*" which was about to revolutionize the approach to Natural Language Processing. The paper addresses the problem of translation which, until that time, was performed using Recurrent Neural Networks (RNNs). RNNs' approach to translation was sequential, meaning that each phrase would be split into single tokens (words) composing the phrase. Each token would be taken singularly, converted into a vector that represented the token and, through an encoder (a function performing operations between vectors) it would be linked to the previous token and the meaning of all words in that sentence would be incorporated into a numerical vector. After having performed this operation for all tokens composing the phrase, the final vector would go through a decoding process which would eventually give as output the sentence translated into a different language.

[23] Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] Available at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

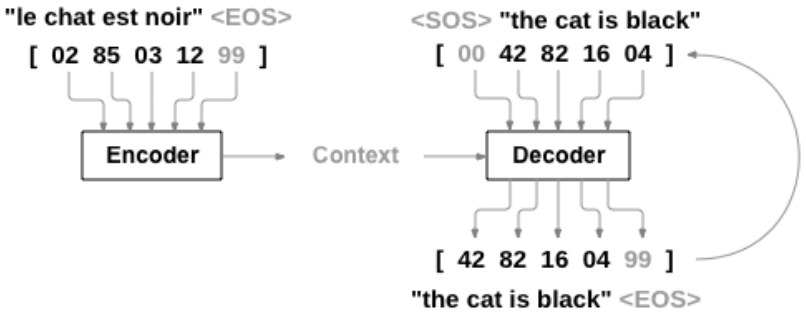


Figure 1: Visual explanation of the RNN process for sentence translation (https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)

However, a relevant problem became evident: the decoding algorithm would only "remember" which was the last word that was translated, and it lacked a general understanding of the global context. This problem is particularly evident in Winograd Schemas[24]: a Winograd schema is a pair of sentences that only differ in one (or two) words and that contain an ambiguity that resolves in two opposite interpretations of the sentence based on which word is chosen. To correctly translate Winograd Schemas a general understanding of the global meaning of the sentence is required. For instance, "*The trophy doesn't fit into my suitcase because it is too large/small*": based on the final word for a human being it is possible to understand if *it* refers to the trophy or the suitcase, hence the translation into French, for example, will be completely different because suitcase and trophy have different genders in French. Therefore, here are the two possible translations in the two cases:

- The trophy doesn't fit into my suitcase because it is too small → Le trophée n'entre pas dans la valise parce qu'elle est trop petite
- The trophy doesn't fit into my suitcase because it is too big → Le trophée n'entre pas dans la valise parce qu'il est trop grand

RNNs did not obtain good results when facing these problems, but the new Transformers' architecture did. As described in Vaswani et al.'s paper, at first a sentence is tokenized, and tokens are transformed into vectors through provided dictionaries that contain what are called the embeddings. Words are not randomly assigned to vectors, and the embedding is itself a great achievement because the values in the vectors carry information about the meaning of the token. Tokens are linked to specific vectors so that mathematical operations between vectors make sense also between the corresponding words. For example, assigning a numerical vector **k** to the word "king" and a numerical vector **w** to the word "woman" and the vector **m** to "man" it is possible to obtain vector **q** as a mathematical result of **q** = **k** + **w** - **m** or, semantically = "king" + "woman" – "man" = "queen".

[24] https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html

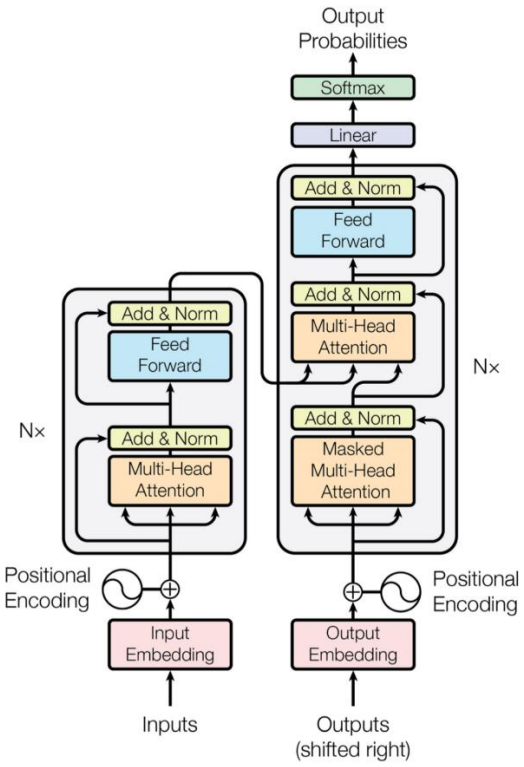


Figure 2: Transformer's architecture from "Attention Is All You Need", Vaswani et al. (2017)

Once this embedding operation is performed, all tokens are in vectorial format, hence it is possible to perform mathematical operations between them. The reason why these operations are performed on vectors is to evaluate the link between each word of the sentence with every other word in the sentence and, therefore, understand the global meaning. As it is possible to observe in Figure 2, there are two parts of the architecture: one for the Inputs and the other one for the Outputs. They both are performed in parallel. At this point of the explanation it is important to introduce the concepts of key (K), value (V) and query (Q): in a sentence a key is a token and a value is the value that the token acquires. For instance, if referring to a person, possible keys can be Name, Height, Age and the values are respectively Alex, 1.80 m, 23. A key and a value can be the same. A query is the specific key the algorithm is currently interested in. Subsequently it will be shown how these are fundamental in the algorithm.

As the sentence gets fed into the architecture, a positional embedding vector is used to remember the place of each token in the sentence. Afterwards, in the three "Multi-Head Attention" blocks the most important parts of the calculations are performed. Indeed, Attention is intended by the authors as a function that "can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, and output are all vectors." It essentially consists of a scalar product between the query vector and the keys vector in order to identify which is the current token which the algorithm is working on to obtain a translation. Once identified the key on which the algorithm is focusing, this is multiplied by the values vector to obtain the value of the token. Attention is computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q, K and V are query, keys and values vectors, $\sqrt{d_k}$ is a normalization factor and the softmax function is used to rescale and select those entries in the scalar product between Q and K that are the biggest, so that the algorithm focuses on those to obtain the values. More details on the architecture and the functioning of Transformers can be found at the following link[25] or by watching these videos[26,27,28]. In the video referenced by number [27] one of the authors of the paper explains the architecture. In the Appendix it is also possible to find an example of the Attention mechanism, Figure A1.

Transformers can be used in multiple areas of Natural Language Processing and one of these is sentiment analysis. In particular, the BERT (Bidirectional Encoder Representations from Transformer) language model represents a cutting hedge model for sentiment analysis. As explained in the paper "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" by Devlin et al. (2018)[29] the model "is designed to pretrain deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers". Specifically, the authors find that unidirectional sentiment analysis models tend to have a worse performance due to the fact that they do not understand the general context. Instead, Devlin et al. manage to train a transformer that can better picture the overall sentiment expressed thanks to two techniques they apply during the training process of the model. The first technique consists of masking or randomly substituting 15 % of all words in the text used for pre-training BERT. In 80% of the cases, these words are substituted with a [MASK] token, in another 10% of the cases they are not substituted and in the remaining 10% of the cases tokens are substituted with other random tokens. In this way, the model learns not to rely on the single token, but to capture the general meaning by observing the words on its left and on its right. A second technique used during the pre-training process is the Next Sentence Prediction (NSP): the corpus from which the model is learning is shuffled so that in 50% of the cases the sentence following is the actual next sentence also in the original document, while in the remaining cases it is not. This trains the model to understand sentence relationships. A fine-tuning of all the 340 million parameters (for BERT large) concludes the training and leads to a model that outperforms pre-existing ones.

It is now clear why for this research it will be used a sentiment analysis algorithm based on BERT, therefore on Transformers, instead of a lexicon-based model. Indeed, not only these models have better performances, but they also require minor pre-processing and no pre-training since they are already pre-trained on specific

[25] https://towardsdatascience.com/transformers-89034557de14
[26] https://www.youtube.com/watch?v=iDulhoQ2pro&t=1333s
[27] https://www.youtube.com/watch?v=rBCqOTEfxvg&t=1580s
[28] https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/bert-encoder
[29] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805.

datasets. For BERT, different versions exist and they are based on the dataset on which it has been pre-trained and fine-tuned: FinBERT is trained specifically on financial texts and is better suited for financial sentiment analysis, whereas BERTweet is trained on a dataset of English tweets. BERTweet is considered more appropriate for this research, given the nature of the dataset, composed of generic tweets.

BERTweet[30] is trained on a corpus made of 850M English tweets, or the equivalent of 16 billion tokens. It contains tweets ranging from 01/2012 until 08/2019 and 5M tweets about COVID-19 from 01/2020 until 03/2020. The architecture is the same as $BERT_{Base}$ described by Devlin et al. (2019) and, compared to many other models, it has the best performance. It is made available to end users in a Python package called Pysentimiento[31,32] and it is the model that was chosen for this research in the sentiment analysis phase, described in the section "Data Analysis". BERTweet assigns labels to the tweets based on the sentiment they express, which can be "POS" for positive sentiment tweets, "NEG" for negative sentiment or "NEU" if the tweet expresses neutral sentiment. Moreover, BERTweet also classifies the feeling expressed in a tweet. The possible emotions are "joy", "anger", "fear", "disgust", "sadness", "surprise" or "others" and these are assigned based on the words used in the tweets.

[30] Nguyen, D., Vu, T. and Nguyen, A. (2020). *BERTweet: A pre-trained language model for English Tweets*. [online] pp.9–14. Available at: https://aclanthology.org/2020.emnlp-demos.2.pdf.
[31] https://huggingface.co/pysentimiento
[32] Perez, J.M., Giudici, J.C. and Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. [online] Available at: https://arxiv.org/pdf/2106.09462.pdf.

3. Research Question & Methodology

Inspired by the research previously conducted, especially by the papers "*Twitter mood predicts the stock market*" by Bollen et al. (2011) and "*The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*" by Oliveira, Cortez, and Areal (2017), this master thesis research will investigate the possibility of predicting the price movements of some financial securities. In particular, the research will focus and deepen the current knowledge on whether it is possible to use sentiment analysis applied on tweets in order to predict a certain financial security's market movements. As a result, the research will ultimately answer to the question "Can the return of a portfolio based on the sectorial ETFs be optimized during the COVID-19 pandemic by using Twitter sentiment analysis to predict stock market fluctuations?".

As discussed in the Literature Review, the topic of using sentiment analysis from content shared on the internet (i.e. on blogs, news websites or social networks) has attracted numerous researchers' interest. However, only lately with the improvement of Natural Language Processing algorithms such as the described Transformers algorithms, researchers were able to study larger amount of data. Such big volumes are necessary to find a shared trend, given the noisy characteristic of social media.

Moreover, the COVID-19 pandemic represents an unprecedented shock to the markets, in a very limited period. This makes the reaction from the population extremely concentrated, with very few elements of disturbance. Therefore, this created some good conditions to study the markets response to external news and the correlation between the price movements and the population sentiment.

In the following steps of the research, firstly the dataset will be created. With a Twitter Research Developer account, about 1.4 million tweets are downloaded. These are specific for the topic COVID-19 and posted from 23rd February 2020 until 31st May 2020 in the US in English, as specified in the following Data section. Also, the 30 minutes intraday market data on specific ETFs is downloaded from Kibot website. These, as well, are specific for the time period analysed.

Subsequently, the research focuses on a first process of data exploration to have a better understanding of the dataset containing the tweets. This process is followed by a process of data cleaning and analysis that ultimately leads to the sentiment analysis of the tweets. The sentiment expressed by each tweet is aggregated in 30-minutes periods to obtain the total sentiment expressed during that timeframe, however always keeping a record of the single number of positive, negative and neutral tweets posted. Also, the financial ETFs data are analysed to obtain the price variation in the 30-minutes timeframe. Then, market data and sentiment data are linked for each 30 minutes period during which the markets were open, in pre-market or afterhours.

Once all these data points have been analysed, classification models such as the k-Nearest Neighbor and the Support Vector Machine are trained and tested to classify data. The two possible classifications are "1" if the security market price is predicted to increase in the 30-minutes timeframe, "0" otherwise. Training and validation of the models are performed on a dataset ranging from 23rd February 2020 until 30th April 2020, while testing is performed on May 2020 data. Finally, the predictions of the best performing model are used to guide investment decisions of a portfolio. If the security is predicted to be increasing in price in the next 30 minutes, then the security is bought at the beginning of the period, if the price is forecasted to be decreasing, then it is sold. The performance of this portfolio is compared to the performance of two other portfolios adopting different strategies in order to assess whether an investment strategy based on insights gained from Twitter sentiment analysis is effective in optimizing the return of a portfolio.

4. Data

4.1 Twitter

Twitter Inc. has become over the years one of the most important social media. With **166 million** Monetizable Daily Active Usage (mDAU) worldwide in the first quarter of 2020, +9% QoQ, and **186 million** mDAU in the second quarter of 2020 worldwide, in the United States there were 33 million in the first quarter (+8% QoQ) and 36 million in 2Q20 (+7% QoQ). It is assumed[33] that every second 6'000 tweets are sent worldwide, therefore about 500 million tweets daily and 200 billion tweets a year.

It is clear that Twitter constitutes an incredible pool of data that can be analysed to comprehend the sentiment of the population with regards to certain topics. Twitter gives the possibility to researchers to create an Academic Developer Account[34] which allows for the download of up to 10 million historical tweets per month. The full archive search allows retrieval of tweets from March 2006. Access is granted for free to academics and students and uses a Twitter API v2 adapted for Python. All times on Twitter are expressed in UTC, hence the Eastern US coast is in the ET time zone which corresponds to UTC – 5 and the Western US coast is in the PT time zone corresponding to UTC – 8. It is possible to code in Python a query so that only specific tweets are retrieved. In this research case, the focus is on tweets from 23rd February 2020 until 31st May 2020, twitted from the United States in English and on COVID-19-related tweets. To specify this, the query specifies that the tweets need to contain at least one of the following words, which are identified as synonyms of COVID-19 or strictly related: "covid", "covid-19", "corona", "coronavirus", "pandemic", "SARS-CoV-2" or "epidemic". Tweets need not to be retweets because they would not add any interesting insight.

Critically discussing the parameters selected for the tweets' retrieval, the choice of selecting only the US as country and English as a language reduce the dataset and could cause some bias. Indeed, the choice of the language might exclude some American citizens posting on the social media from the US but in other languages, such as Spanish or Chinese. These tweets are also part of the sentiment felt in the nation and might influence the sentiment of the markets or of minorities speaking that language. Also, restricting the analysis to US tweets, excludes all tweets that come from different countries, but which are somehow related to the American economy and might influence it. It also implies that it is possible to retrieve only those tweets for which the localization is turned on, so some users who have not given access to their location will not be analysed. Regarding the choice made on the keywords, it is evident that these selection rules applied on the text may exclude many tweets related to the topic, which, however, do not contain these keywords. A possibility would be to include slang definitions of COVID-19, such as the word adopted by Donald Trump to negatively define COVID as "kung flu". However, these would not guarantee optimal results as any tweet

[33] https://financesonline.com/number-of-twitter-users/
[34] https://developer.twitter.com/en/products/twitter-api/academic-research

containing the word "flu" could be identified and taken for a definition of COVID-19 even though it may refer to a normal flu. To conclude, the Twitter API v2 has a limited research parameters length, which is already reached with the previous research rules.

Overall, the rules defined above guarantee a filter for tweets and reduction of the possible noise that would be generated if all the tweets were downloaded indiscriminately. In total, 1,429,450 tweets were downloaded, spanning over the period from 23rd February 2020 until 31st May 2020. The distribution is shown in the graph below, Figure 3. It is clear that the peak is reached on 12th March 2020, when more than 52,000 tweets were twitted with the characteristics described above.

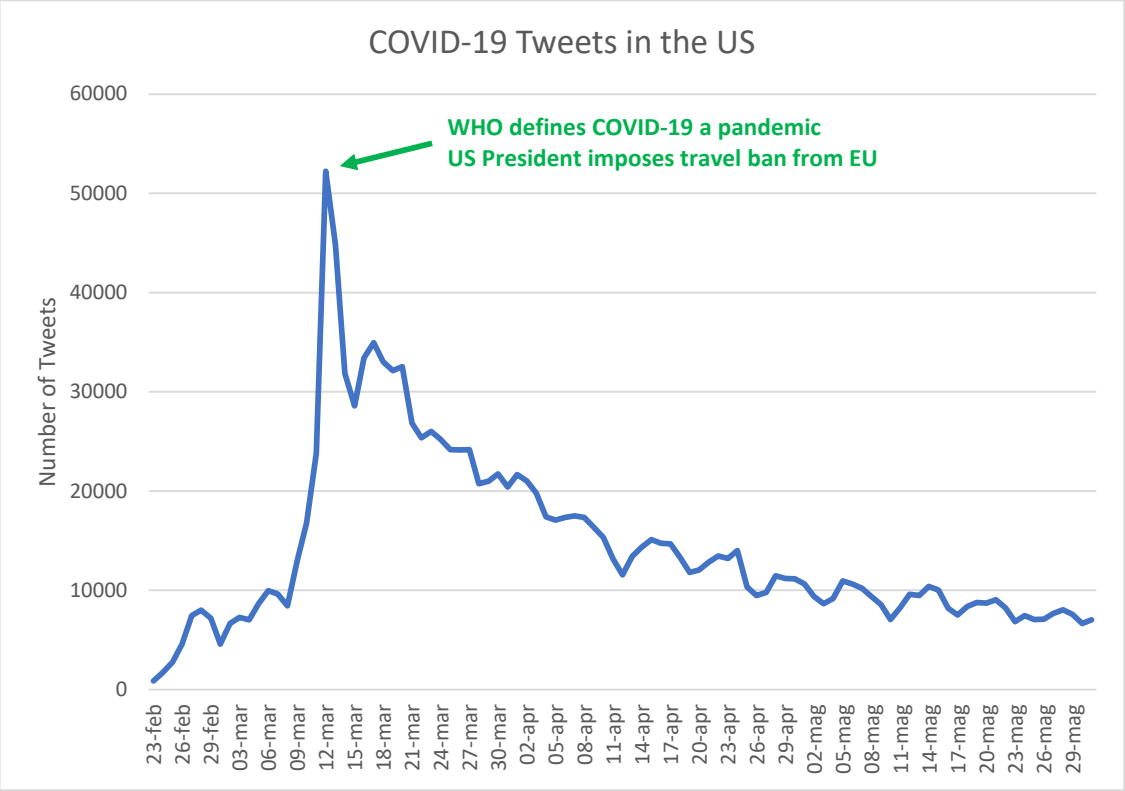


Figure 3: Number of daily tweets posted with the characteristics defined. Overall, they sum to about 1.43 million tweets from 23rd February 2020 until 31st May 2020. The peak corresponds to the 12th March 2020, the day after the WHO declared COVID-19 a pandemic and President Donald Trump imposed a travel ban for flights from the EU.

This day also corresponds to the biggest drop (-10%) of the DJIA index since Black Monday in 1987. The extreme surge in number of tweets follows the declarations of the World Health Organization on 11th March 2020, when COVID-19 was declared a pandemic. Also, late on 11th March 2020 US President Donald Trump had announced a 30-days travel ban from the European Union. On 12th March 2020 all major market indexes suffered heavy losses: the DJIA lost -10%, the S&P500 plunged 9%[35].

[35] https://www.cnbc.com/2020/03/12/stock-market-today-live.html

The information downloaded regarding tweets include: the date and time expressed in UTC time zone, the location which includes the city and state, the country, the language, the number of likes, retweets and comments the tweet received, the text of the tweet.

created_at	location	country_code	lang	like_count	quote_count	reply_count	retweet_count	text
2020-02-24 05:45:01	Utah, USA	US	en	1	0	0	0	@AP @HHSGov @CDCgov @WHO#coronavirusThe judge is denying mo
2020-02-24 05:45:10	Aptos, CA	US	en	1	0	1	0	@DoodlesTrks With everything in the news, it's no wonder you're fright
2020-02-24 05:46:36	Winter Park, FL	US	en	1	0	0	0	Hard to explain. Were regimes in China & Iran somehow working t
2020-02-24 05:48:11	Harvey, LA	US	en	0	0	1	0	@biY3aiW59XumE7Z @COVID_19NEWS Wait, did that actually happen?
2020-02-24 05:48:27	Fairview, NJ	US	en	65	9	4	22	Wait. Italy it is now. #corona #coronavirus #coronavirusoutbreak https:
2020-02-24 05:51:27	Kaanapali, HI	US	en	0	0	1	0	@biY3aiW59XumE7Z @COVID_19NEWS I hate when I'm even more cor
2020-02-24 05:51:43	Enterprise, NV	US	en	0	0	0	0	"These people are at Travis Air Force Base right now — why can't they s
2020-02-24 05:54:39	Lakeland Heights, TX	US	en	1	0	0	0	Grief is hard. I lost my bestfriend almost 20 years ago because her husb
2020-02-24 05:56:07	Castlewood, CO	US	en	3	0	0	1	#SouthKorea becomes newest front in shifting #COVID19 #coronavirus
2020-02-24 05:56:13	Ludlow, MA	US	en	0	0	0	0	Hey Twitter land been real busy no recent tweets i might have been mis
2020-02-24 05:59:21	Delaware, USA	US	en	4	0	0	0	Show without public #armani #milano #coronavirus https://t.co/BX4k0
2020-02-24 06:06:06	Lakewood, OH	US	en	1	0	0	1	Corona virus looks like https://t.co/pKFMkpnvjc
2020-02-24 06:06:57	Pleasanton, CA	US	en	0	0	0	0	I'm cryin lml#coronavirus #amh #maskoff #maskon @ New York, New Yo
2020-02-24 06:06:57	Silver Spring, MD	US	en	0	0	1	0	China having problems to explain leak of coronavirus from their laborat
2020-02-24 06:10:37	Berkeley, CA	US	en	0	0	0	0	Corona virus and supply chains.
2020-02-24 06:11:08	Home Gardens, CA	US	en	1	0	1	1	Happening in China thy are Killing people what the Corona Virus so for
2020-02-24 06:12:06	Itasca, IL	US	en	0	0	0	0	"Russian authorities target Chinese nationals with raids and facial recog

Figure 4: The information downloaded from the Twitter Developer Account

4.2 Exchange Traded Funds (ETF)

The financial crisis generated by the COVID-19 pandemic did not hit all companies in the same way. Indeed, some industries even benefitted from the viral outbreak e.g. the healthcare industry because of the increased expenditure in medical equipment and pharmaceuticals as well as the enormous amount of funds received for research on COVID-19 vaccines. Another example is the Tech industry which benefitted from the "work-from-home" policies and the Online Retailers that saw a surge in orders online because people would not go out to buy from physical shops. On the other hand, the Entertainment industry suffered from the sudden stop on tourism, restaurants and hotels, the Oil and Car industry from the lower consumes. These are just examples showing how different sectors experienced very different impacts. Therefore, in this research the analysis will be conducted on different industries.

The industry performance will be replicated by ETFs. Exchange-Traded Funds operate by replicating the performance of an index, sector, commodity or other assets: they are exchanged on the market and are baskets containing various underlying which could be, for instance, the companies of the S&P500 Index. In the case examined in this research, the focus will be on the Industrial Select Sector SPDR Fund (XLI) representing the industrial sector, the Financial Select Sector SPDR Fund (XLF), the Technology Select Sector SPDR Fund (XLK) and the Health Care Select Sector SPDR Fund (XLV). These ETFs were chosen given the exposure of these sectors to the pandemic. Since the influence of sentiment expressed through Twitter can be presumed to be short-lasting, therefore intraday data are needed on the price of the ETFs. This means having access to very detailed data with a granularity that is not available for free on the internet. The data about the chosen ETFs regarding, among all, the price and the volume with a 30-minutes frequency was

downloaded from Kibot. Data includes pre-market (8:00-9:30 a.m. ET), regular (9:30 a.m. - 4:00 p.m. ET) and after-market (4:00-6:30 p.m. ET) sessions and all times are expressed in ET time zone.

5. Data Analysis

5.1 Exploratory Analysis: WordCloud and Clustering

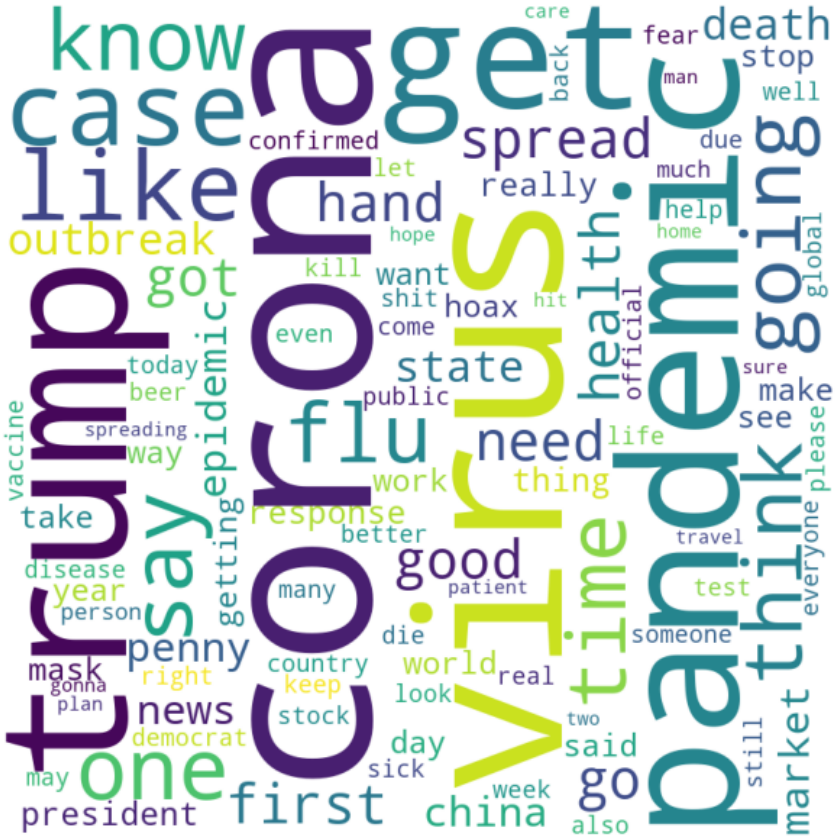


Figure 5: WordCloud representing the 100 most significant words contained in a subset of the tweets dataset made of 10,000 tweets. The importance of each word is calculated by using the TF-IDF technique.

The process of data analysis begins with the tweets dataset. The most critical part of dealing with textual data is the complexity it represents and the rare structure of the texts. Tweets can contain text of maximum 280 characters, but these may also contain links, emojis, punctuation, tags, peoples' names. This variability complicates the analysis and requires a first exploratory analysis of the dataset. To do so, only a small random subset of the dataset of 10,000 tweets has been selected. This was necessary due to the extreme computational power required to perform textual cleaning and in order to being able to obtain representative results. The 10,000 tweets sample is randomly chosen from tweets posted from 23rd February 2020 until end of March 2020. To be able to explore the dataset, a pre-processing of the textual data in the tweets is necessary, which enables a WordCloud representation of the 100 most representative words along with a 10 clusters classification.

Pre-processing texts requires a first step of lower case conversion, in which all letters are converted into lower case to have a homogeneous dataset. The next step is tokenization: it consists of splitting sentences

into single words, so the result is a bag of words, a dataset of single words. These are then going through a process called lemmatization which consists of removing prefixes and suffixes from the words in order to obtain the root of the word. This procedure avoids counting as different words plurals, for example, or the same verb conjugated. Words are also filtered by using an English dictionary, so that all unknown words, emojis, web links and punctuation are filtered out. The last step consists of filtering stop words such as articles, connecting words such as “so”, “then”, “and” etc. which are meaningless for this analysis and in understanding the content of tweets. Also, specific words are filtered out: in this case words such as pandemic, COVID-19, numbers will be extremely present, but their presence is trivial and would not add any interesting information. After this process is completed, there will be a subset of meaningful words.

The subset of words aforementioned, can be represented in a WordCloud, representing all the most significant words appearing in the tweets. Importance is represented by the dimensions of the word in Figure 5 and is assigned to words based on their TF-IDF score.

The TF-IDF method is a statistical measure that consists of computing the importance of a word in a set of documents. TF-IDF stands for Term Frequency - Inverse Document Frequency and it is formulated in a way that the term frequency of a word within the document is calculated and multiplied by the logarithm of the ratio between the number of documents in the set and the number of documents in which this word appears. The rationale behind this formula is that a word grows in importance if it is frequent in a document, and it is not present in all documents. Indeed, this means that a document discusses specifically a topic which is represented by the frequent word in that document, but it is not present in the other documents, so it is not a trivial word. The TF-IDF creates a matrix of importance for all the words in the basket and, from this, it is possible to create a WordCloud plot (Figure 5).

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|} \qquad idf_i = \log_{10} \frac{|D|}{|\{ d : i \in d \}|} \qquad tf - idf_{i,j} = tf_{i,j} * idf_i$$

Part of the exploratory analysis is the clustering process. The subset analysed is the pre-processed 10,000 tweets set, a clustering algorithm is run sorting words in a different number of clusters from 1 to 20 and computing what is called the sum of squared errors (SSE) (Figure A2). The SSE is the sum of all squared distances between the centroid of each cluster and all the points in the cluster. The computation of the SSE is used to determine the best number of clusters to adopt for the final grouping of the words. In this case, the number of clusters chosen is 10, and for each of them the algorithm shows the 10 words most representative of the content of each cluster. As shown by the results, some clusters are extremely defined in the sub-topic of tweets contained in them: cluster 4 contains words that can be linked to tweets that had a “reporting” function about the number of new cases and deaths. Cluster 6 contains tweets where people express their opinions considering COVID-19 as a "hoax" as it was at the beginning defined by President

Trump. Cluster 0, although it may contain very different tweets, from the most representative words it can be deducted that the authors were encouraging people to “stay at home”.

Cluster Number	Number of Tweets in the Cluster	Most Representative Words
Cluster 0	4630	'need', 'one', 'test', 'due', 'take', 'please', 'positive', 'home', 'stay', 'say'
Cluster 1	621	'corona', 'got', 'shit', 'fuck', 'really', 'beer', 'time', 'need', 'man', 'one'
Cluster 2	616	'virus', 'corona', 'get', 'know', 'got', 'really', 'shit', 'flu', 'need', 'stop'
Cluster 3	431	'like', 'feel', 'look', 'pandemic', 'corona', 'got', 'one', 'know', 'time', 'day'
Cluster 4	302	'case', 'confirmed', 'county', 'first', 'state', 'death', 'total', 'positive', 'number', 'report'
Cluster 5	947	'day', 'time', 'know', 'going', 'week', 'paper', 'last', 'toilet', 'home', 'work'
Cluster 6	477	'trump', 'president', 'pandemic', 'response', 'administration', 'say', 'said', 'crisis', 'hoax', 'test'
Cluster 7	261	'right', 'thing', 'pandemic', 'corona', 'like', 'get', 'need', 'one', 'going', 'home'
Cluster 8	537	'get', 'corona', 'pandemic', 'tested', 'sick', 'hope', 'need', 'let', 'going', 'want'
Cluster 9	972	'pandemic', 'global', 'even', 'world', 'going', 'need', 'time', 'one', 'still', 'make'

5.2 BERTweet: Sentiment Analysis

This preliminary analysis highlighted the structure of tweets and what can be expected during the subsequent analysis. However, it is essential to specify two decisions taken for the following part of data analysis. The first resides in the fact that this research focuses on all tweets posted by users with any kind of background and that may or may not express an opinion linked to the financial markets. Indeed, one of the challenges of this research is to examine whether it is possible to extract financially insightful information from any kind of tweet. This is the reason why it has not been chosen to select users that only share news or market related topics, or only the most influential ones such as the New York Times' Twitter account, for instance. Therefore, the scenario taken into consideration in this research is the most generalized one, where there is no previous research on which accounts are the most influential, nor on the truth of the comments posted. Out of this scenario, the objective is to find out if any relevant financial insight can be extracted. To reach this goal more than 1.4 million tweets need to be analysed by using BERTweet. BERTweet is optimized for textual analysis, with textual pre-processing included in the algorithm, and thus the use of the transformer algorithm will be preferred to manually implementing the computationally expensive pre-processing method and the less accurate lexicon-based algorithms.

After the exploratory analysis, it is necessary to analyse the tweets to characterize the sentiment they express. Tweets are fed to BERTweet via the Python package *pysentimiento*, which, because of the algorithm explained in the Literature Review section, establishes if a tweet is positive (POS), negative (NEG), or neutral (NEU) and provides the probability of such result. As a result, the following is the statistics representing the sentiment over the considered period. From the table below it is possible to observe a higher majority of negative tweets over positive tweets during the whole time period. More detailed graphs can be found in the Appendix, figures A4 and A5.

Period	Positive Tweets	Neutral Tweets	Negative Tweets	Total Tweets
23rd Feb – 31st Mar	105'836 (14.5%)	293'721 (40.1%)	331'795 (45.4%)	731'352
1st Apr – 30th Apr	79'135 (18.3%)	161'761 (37.5%)	190'835 (44.2%)	431'731
1st May – 31st May	49'030 (18.4%)	92'556 (34.7%)	124'779 (46.8%)	266'365

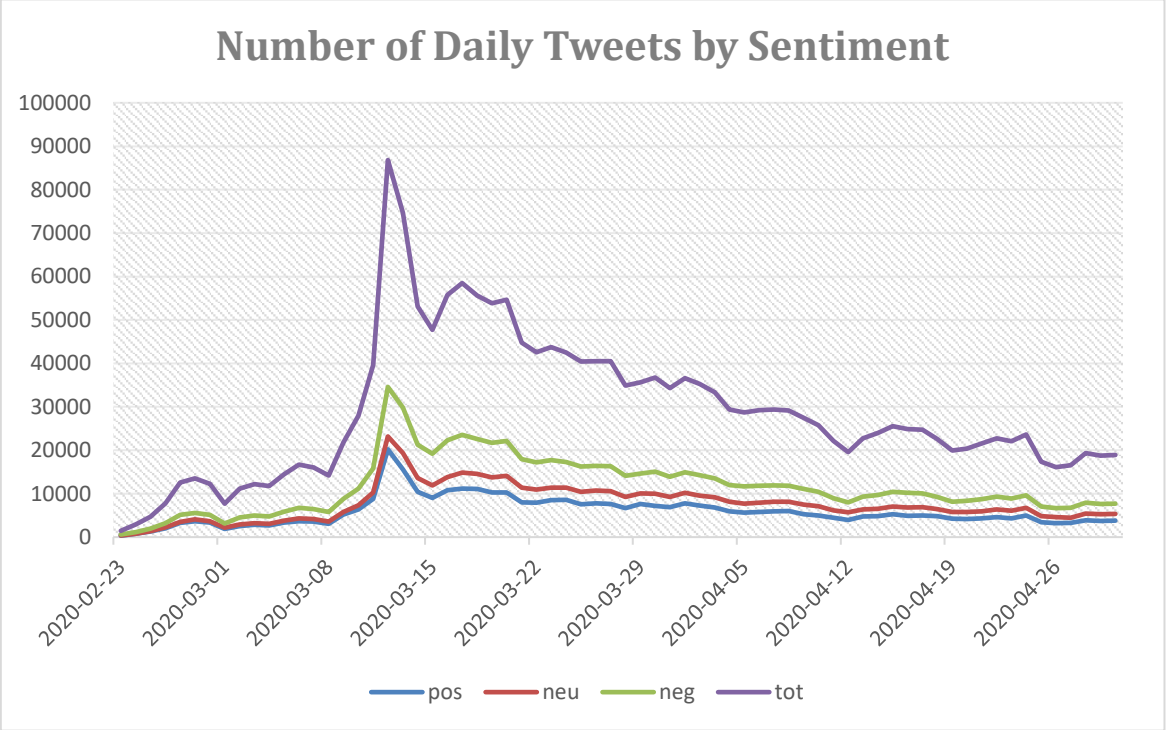


Figure 6: The number of positive, neutral and negative tweets in the train dataset.

Once each tweet has been labelled, the algorithm works on identifying the sentiment of each time period during which the stock market in the U.S. was open. To reach this goal, the first step is to uniform the times by shifting the financial market data on ETFs from the E.T. time zone to the UTC time zone. It is to be noted that on 8th March 2020 the Daylight-Saving Time began, so from 23rd February 2020 until 7th March 2020 the times indicated in the financial data are moved forward by 5 hours, and from 8th March 2020 until 31st May 2020 the times are moved forward by 4 hours. At this point, the goal is to characterize the sentiment that was expressed on Twitter in the time period expressed by the financial data and link the two information about the sentiment and about the ETFs' performance. This procedure will allow for a model to find a correlation between the markets fluctuations and the sentiment fluctuations on Twitter. Since the granularity

of data available is 30 minutes, then the sentiment will be characterized every 30 minutes for all the periods of time during which the markets were in pre-market, open and after-hours (from 8:30 E.T. until 16:30 E.T.).

To identify the sentiment, the algorithm proceeds with an aggregation of the sentiments expressed by the tweets: for every period of 30 minutes corresponding to the same ETFs date and time period, the algorithm counts how many positive, negative and neutral tweets have been posted, as well as counting the total number of tweets posted. However, to remove some uncertainty over the correct sentiment label of a tweet, the algorithm only considers those labels with a probability of being correct higher than 70% which was assigned by BERTweet during the labelling process. The same counting procedure is applied to the feeling expressed. The result is an aggregation for every 30 minutes with the performance of the market linked to the sentiment expressed during the same timeframe. The number of positive, negative, and neutral tweets are expressed as percentages over the number of total tweets and so is the number for each feeling expressed in the tweets. Also the ETF's price change is expressed in percentage compared to the price at the end of the previous 30 minutes interval. Figure 7 shows an example.

		date_time_UTC	open	close	volume	change	neg	neg/tot	pos	pos/neg	neu	pos/tot	tot
1													
2	**0**	2020-02-24 14:00:00	78,38	78,37	13422	-0,01276	28	0,528302	2	7,1%	12	3,8%	53
3	**1**	2020-02-24 14:30:00	78,28	78,54	2500621	0,332141	29	0,453125	2	6,9%	21	3,1%	64
4	**2**	2020-02-24 15:00:00	78,53	78,34	1548106	-0,24195	35	0,472973	0	0,0%	26	0,0%	74
5	**3**	2020-02-24 15:30:00	78,33	78,49	1678638	0,204264	20	0,47619	2	10,0%	15	4,8%	42

Figure 7: Example of the aggregated data: to each 30 minutes time period corresponds a total number of positive, negative and neutral tweets posted, as well as the percentage change in market price. The ratios **positive/total**, **negative/total**, **positive/negative** are also computed.

5.3 Choice of Models for Market Movement Prediction

5.3.1 Linear Regression

With this data, a first possibility is to predict the exact price variation of the asset by implementing a linear regression model. The use of linear regression has the purpose to find the exact percentage of increase or decrease in the market following a given change in the sentiment. Therefore, the first approach aims at forecasting the ETF's movement precisely. It is possible to consider the percentages of positive, negative tweets and of total tweets as independent variables, since the model does not include the number of neutral tweets, which would cause a dependency. Therefore, it is possible to perform a linear regression using these three independent variables, and the percentage change in the ETF's price as dependent variable. After performing the analysis, some valuation techniques are applied, especially regarding the R-squared values. Based on this simple metric, it is possible to conclude that the linear regression model does not perform well, since R-squared values are extremely low and close to 0. The main possible reasons for this result are that:

- the problem is not linear. Indeed, the relation between these variables might be better represented by non-linear functions.
- the forecast of exact market movements is an extremely complex problem, which requires more complex models.

5.3.2 Support Vector Machine and k-Nearest Neighbour for Directional Prediction

The second possible use of the dataset is to employ the data to make a directional forecasting of the market's movements. This will become the main goal of this research. For directional forecasting it is meant only the direction in which the market will move: increase or decrease, over the 30 minutes period. Therefore, the problem becomes a classification problem. With the slight change of approach, also the dataset needs a modification in the way data are represented. For instance, the dependent variable that indicated the percentual change in market price only needs to represent if the movement has been an increase in market price or a decrease. Therefore, for every 30 minutes timeframe, if the price of the ETF increased compared to the previous period, then the label acquires the value 1, if it decreased, then the label is 0. The same is done for all the other independent variables that will be used as inputs for the models: if the fraction of positive tweets over the total count of tweets has increased compared to the previous period, then it acquires the value 1, and 0 otherwise. Hence, all features and the target variable too are dummy variables with values (0,1) apart from the total number of tweets posted in the period, which is rescaled to 1 as *(total number of tweets in that period)/(maximum total number of tweets in a 30 minutes period)*. An example can be found below in Figure 8.

1_up	neg/tot_1up	pos/neg_1up	pos/tot_1up	tot_scaled	joy_1up	anger_1up	fear_1up	disgust_1up	sadness_1up	surprise_1up	others_1up
0	1	1	1	0,000643	1	1	1	1	1	1	1
1	0	0	0	0,000776	0	0	0	1	0	0	1
0	1	0	0	0,000898	0	1	1	0	1	1	0
1	1	1	1	0,000510	1	0	0	0	0	0	0

Figure 8: for directional prediction of the stock market movements, the problem is characterized as a classification problem. Variables are characterized as dummy variables, where 1 indicates an increase in the value compared to the previous 30 minutes period. For example, the third line indicates that a decrease (0) in the ETF price "1_up" column during a certain 30 minutes period verified while there was an increase (1) in the "neg/tot_1up" (negative/total tweets) compared to the previous period, a decrease (0) in the positive/negative tweets ratio, a decrease in the positive/total tweets ratio, a total of 0.000898*MAX(Tweets posted in a 30 minutes period) so that the maximum value is normalized to 1. Also, emotions indicate that during that timeframe "joy", "disgust" and "others" decreased (0), "anger" increased (1), as well as "fear", "sadness" and "surprise".

These modifications to the dataset allow classification to be made. Therefore, the objective is to predict whether the ETF price movements will be an increase (1) or a decrease (0). This is a classification problem which can be solved via classification algorithms. Although there may be many different classification models, as seen in the papers described in the Literature Review[36], the best performing ones have been the K-Nearest Neighbour, KNN, algorithm and the Support Vector Machine, SVM, algorithm.

[36] Oliveira, N., Cortez, P. and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, pp.125–144. doi:10.1016/j.eswa.2016.12.036

K-Nearest Neighbours[37] (kNN) is a supervised machine learning algorithm which can be used for both regression and classification. It is based on the concept that a new point that the algorithm is trying to classify is more likely to belong to the same class which the majority of its k (k being an odd natural number) nearest neighbours belong to. Usually, kNN is run multiple times with different values of k, in order to decide which is the best number of neighbours to consider for the best accuracy in the results. The algorithm, therefore, identifies areas based on which it will be able to classify any new datapoint that has not a label yet[38]. This algorithm is apt for linear and non-linear problems. Figure 9 provides an example.

Support Vector Machine[39] (SVM) with linear kernel is a linear model used for regression or classification. If used for classification, it consists of finding the hyperplane that best separates two classes. With reference to the figure below, Figure 10, there are multiple lines separating the two classes. The best one among these can be identified by finding the two points (belonging to two different classes) closest to the line, the distance between the points and the line is called margin, the best hyperplane is the one maximizing that distance. SVM can also be applied to non-linear problems[40].

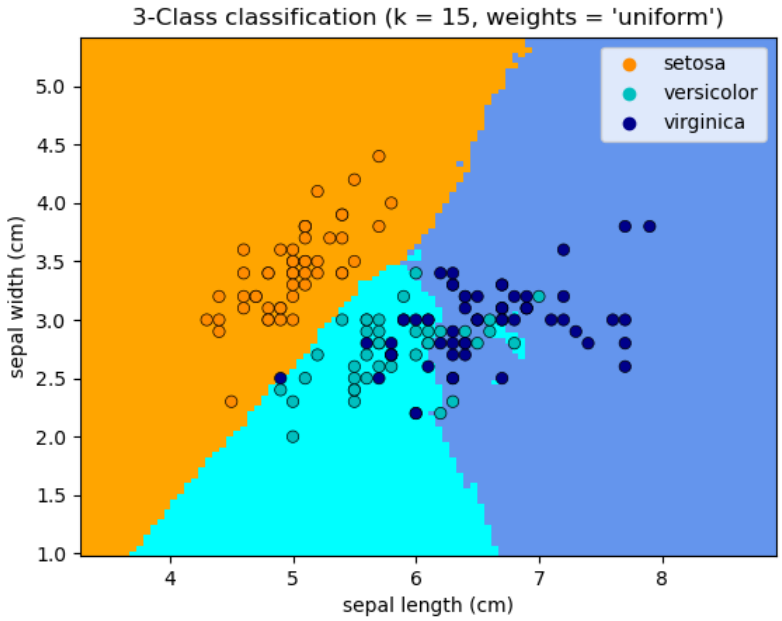


Figure 9: Example of classification using kNN with k = 15 from Scikit Learn[41]

[37] Towards Data Science: k-Nearest Neighbours https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3
[38] Scikit Learn: k-Nearest Neighbours https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[39]Towards Data Science: Support Vector Machines. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989#:~:text=SVM%20or%20Support%20Vector%20Machine,separates%20the%20data%20into%20classes.
[40] Scikit Learn: Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html
[41] https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py

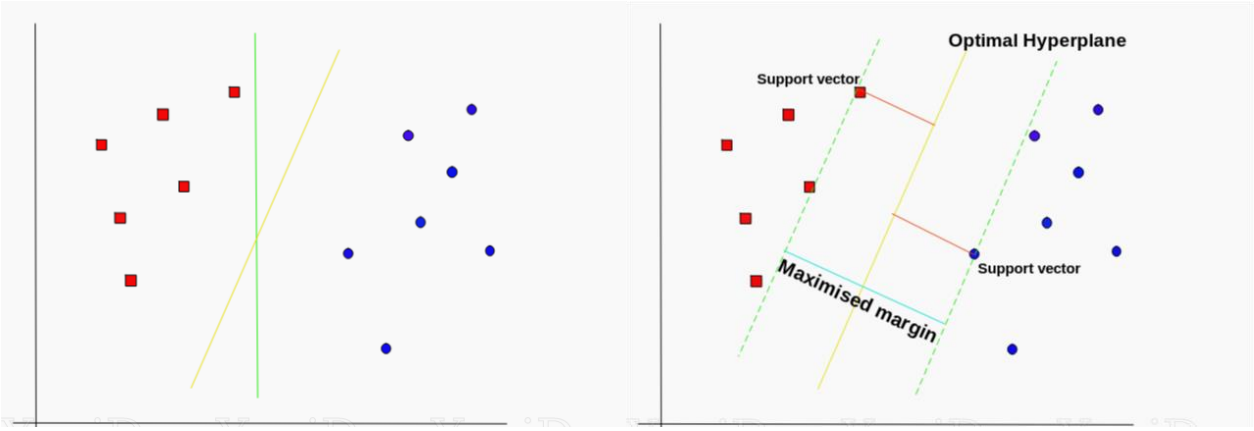


Figure 10: Support Vector Machines are based on the concept of finding an hyperplane that best separates the classes.

In the following part, the two models will be trained, validated and tested and the results compared. The two models rely on supervised training, which means that the algorithm receives the datapoints which are made of different features such as the ratios “positive/total”, “negative/total”, “positive/negative” and “total” and the label which is the change in market price compared to the previous 30 minutes. This label assumes the value 1 if the price has increased and 0 otherwise. The training and validating dataset consist of data from 23rd February 2020 until 30th April 2020, for a total of 715 datapoints. Of these points, 339 are labelled as 0, corresponding to 47.4%, while the remaining 376 are labelled as 1, 52.6%. So, the dataset is balanced. The test dataset is made of data from 1st May 2020 until 31st May 2020, for a total of 298 datapoints. Out of these points, 142 are labelled as 0 (47.7%), the remaining 156 (52.3%) are labelled as 1. The testing dataset is balanced as well.

The training part of the process consists in making the model “learn” the correct parameters to set the hyperplanes in the SVM (or borders in case of a kNN) separating the classes. Then the validation part consists of using a fraction of the training dataset to score the accuracy of the model on data for which the label is known. The training and validation processes are performed with an approach called k-fold cross validation. Specifically, this means that the whole training subset of 715 datapoints is used. A 5-fold cross validation approach is implemented, which means that the whole set is divided into 5 parts, each of those corresponding to 20% of the dataset. Then, the algorithm is trained on the first 4 subsets of the dataset (80%) and validated on the last subset (20%), the accuracy result and the learnt hyperparameters are saved and the model is trained again on the whole dataset, but the fourth 20% subset, which is used for validation. The algorithm saves the accuracy and adjusts the hyperparameters. This process continues until all the subsets have been used once for validation. This allows to train the model on the whole dataset and avoids overfitting. An example of the 5-fold cross validation process can be seen in the Appendix, Figure A3.

There are many features for each datapoint representing the information on the ratios “positive/total”, “negative/total”, “positive/negative”, “total”, “joy/total”, “anger/total”, “fear/total”, “sadness/total”, “disgust/total”, “surprise/total” and “others/total”. Since the importance of each feature for the model is

not known a priori, then the training and validation processes are run on SVM and kNN for all possible combinations of these features. For example, in a training and validation process the algorithm will only take into consideration the features “positive/total”, “negative/total” and “total”, in another process it will consider “positive/negative”, “total”, “positive/total” and “joy/total” and so on, until it considers all possible subsets of combinations. The total number of possible combinations is $2^{Features}$, hence 2048 possible subsets on which the two models are run. This process is aimed at finding the best and most insightful subset of features, to run afterwards a targeted test process with only the selected features. The kNN algorithm also requires the choice of the best number of neighbours (k) to consider for a better accuracy of the model. Therefore, during the training process the kNN model is run with k = {1,3,5,7} for a total of 2048 combinations for each k. The two models are run with the same process described above for different ETFs, which are representing the most affected industries by the COVID-19 pandemic.

Once the results of the train and validation process are available and analysed, the best performing models for SVM and kNN are run on the test set. For SVM the only parameters chosen to perform the trained model on the test set are the features to include. Indeed, the process of running the model on the 2048 subsets highlights the best subset of features to include to have the best accuracy. For the kNN algorithm, in addition to choosing the best features subset, it is also necessary to choose the number of neighbours to use to implement the trained kNN model on the test set. A complete discussion of the results obtained is performed in the following section.

6. Results and Discussion

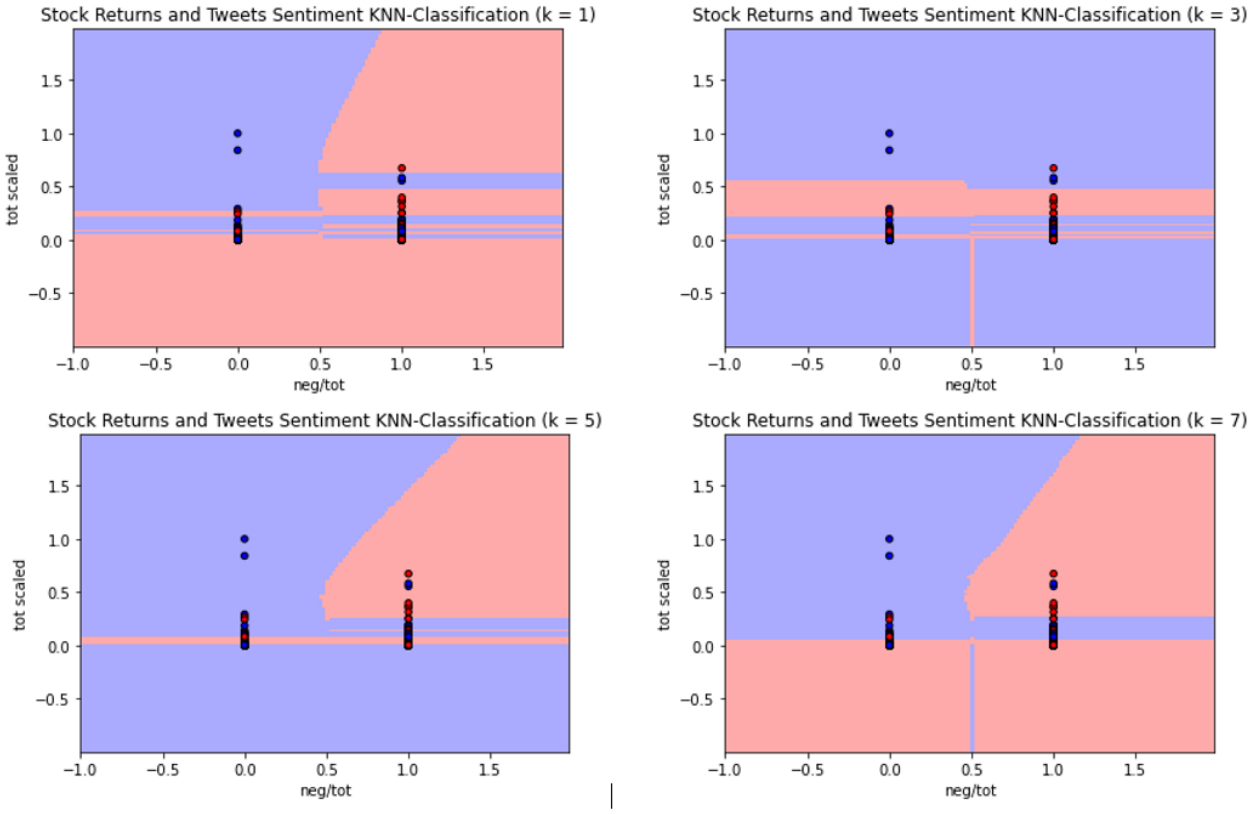


Figure 11: Visual representation of the kNN algorithm with respectively k = 1, 3, 5, 7. These graphs consist of a 2D representation (2 features only). For the research, subsets ranging from only 1 feature to all 11 features have been considered, therefore the above graphs have to be imagined in many more dimensions. As explained below, 4 features (4 dimensions) have proven to give the best results.

The SVM and kNN models are trained on data from 23rd February 2020 until 30th April 2020 and tested on data in the period from 1st May 2020 until 31st May 2020. The chosen ETFs are the ones mentioned before, which are considered to be possibly the most influenced by the pandemic: the industrial XLI, the financial XLF, the healthcare linked XLV and the technological XLK. The training and validation results are summarized in Figure A5 in Appendix . Only the highest accuracy results have been reported for each ETF. If a high accuracy level is reached for any k value in the kNN algorithm, then also the SVM value for the same subset of features is reported, as well as the value for all k values in the kNN, and vice versa.

From a first analysis of the results, it is possible to observe that most of the high accuracy results are obtained for subsets of either 3 or 4 features. The reason behind this finding is that an excessive number of features leads to an overfitting of the model and eventually adds noise to it. Indeed, some features result in not being important for the model and not carrying useful information for the classification. Moreover, not basing the models on too many features is pivotal in avoiding the curse of dimensionality, which would make calculations much harder and prevent models from performing at their best.

A second finding is that the features describing the sentiment such as "positive tweets/total tweets", "negative tweets/total tweets", "total tweets" and "positive tweets/negative tweets" result in being more

insightful for classification. Among the feelings, the most recurrent with high accuracy are "fear", "surprise" and "sadness". In general, in all the highest scoring subsets there is only one feature describing the feeling and most of them describe the sentiment. A possible explanation for this is that labelling the sentiment is easier for BERTweet than labelling the feeling, for which it may not be optimized.

The fact that the best features chosen when analysing the different ETFs may not always be the same is due to the fact that different economic sectors may be influenced by different sentiments and feelings expressed on Twitter. For example, the financial ETF seems to be more dependent on feelings of surprise and fear, while the technological one depends on "fear", "sadness" and "anger" and the healthcare ETF is better classified with "sadness" and "fear" feelings. The industrial sector, on the other hand, seems not to have any strong dependency on feelings, but rather on sentiment.

The best scoring subsets in the validation process are then used to perform the testing procedure on the May data. Overall, the accuracy reached on the test set is only a few percentage points above 50 %. kNN performs best with a value of k = 5 on **XLI (58.25%)** and **XLF (55.51%)** where the value in brackets represents the accuracy reached. SVM overperforms kNN only for **XLV (52.53%)** and **XLK (57.38%)**, and it has the same performance for XLF (55.51%) but on a different subset of features. All the results for the best performing algorithm settings can be found in Appendix Figure A5.

Therefore, the best result of 58.25% accuracy is reached for the industrial index, using the features regarding sentiment "negative tweets/total tweets", "positive tweets/total tweets", "positive tweets/negative tweets" and "total tweets". It is well known that the industrial sector has heavily suffered during the pandemic due to the necessity to stop factories because of the spread of the virus, so it is understandable that the sentiment expressed on Twitter might be related to the performance of the sector. Indeed, this implies that a perceived improvement in the pandemic situation may have impacted the price of the securities in the sector, making it possible to correctly classify the market performance given the Twitter sentiment.

Although the accuracy results obtained may seem negligible, this is not the case for this research. Accuracy results above 58%, as the one obtained for XLI, show that it is possible to extract financially insightful information from the tweets submitted by all the population, without implementing any sort of filtering procedure on the users posting the tweets. Considering the high amount of noise and the difficulty of textual analysis, this represents an important result. This result corroborates the previous findings mentioned in the literature review and suggests that sentiment analysis of social media posts could be used to help making investment decisions. The results obtained do not show a strong overperformance of one of the two chosen models over the other. It is possible to observe that with kNN normally higher accuracy is reached, but this does not apply for all analysed ETFs.

To confirm the findings, the performance of three portfolios will be compared investing in the industrial ETF XLI, since it is the one that led to the best accuracy:

- A portfolio (**portfolio A**) that adopts the investing strategy of buying USD 1M value in the industrial ETF on the 1st May 2020 and keeps it until the 31st May 2020, without reacting to changes in the markets or the news.

- A portfolio (**portfolio B**) that has a budget of USD 1M to invest in the industrial ETF adopting the investing strategy suggested by the predictive algorithm kNN for k=5. Every 30 minutes interval when the market is open, the algorithm predicts whether the security value will increase or decrease in the interval. The output is 0 for a decrease and 1 for an increase. Given a portfolio, in case the algorithm predicts a decrease (0), then the algorithm will sell all the securities in the portfolio because it means that a decrease in the market price is predicted to occur. Oppositely, if an increase is predicted (1) and given an empty portfolio, then the algorithm will buy the equivalent of USD 1M value in XLI, expecting to benefit from a market price increase. In case the portfolio is empty and the algorithm predicts a decrease, then nothing happens. Also, no action is performed in case all the money available has already been invested and another increase (1) is predicted, because in case of subsequent increases in the market price, it is best to keep the security until the next decrease (0) signal to fully benefit from the market price increase and not to incur in transaction fees. On 31st May 2020, after market close, the performance of the portfolio is evaluated.

- A portfolio (**portfolio C**) that implements the previous strategy of buying at increase (1) signals and selling at decreasing (0) signals, but whose predictions do not depend on the classification algorithm. Instead, the signals (0 or 1) will be random with a 50 % probability each for every 30 minutes interval.

Extremely interesting results are found supporting the previous findings and suggesting a correlation between the markets performance and the sentiment expressed on Twitter. Indeed, while portfolio A has a total return of 7.33 % during the period, portfolio C has a return of 7.58 % and **portfolio B overscores with a total return of 15.84 %.** Portfolio A represents the strategy of a passive investor and represents therefore the baseline. Portfolio C represents a random strategy which has a different return every time it is applied, since the investment decision is based on a random variable with 50 % probability of being a sale order and 50 % probability of being a purchase. The fact that the strategy implemented in portfolio B leads to more than twice the return of portfolio A suggests the relevance of this research. Indeed, by being able to correctly classify certain 30 minutes periods allows for an optimized investing strategy based on selling at the peak and “buying the dip”, as it would be said in financial jargon.

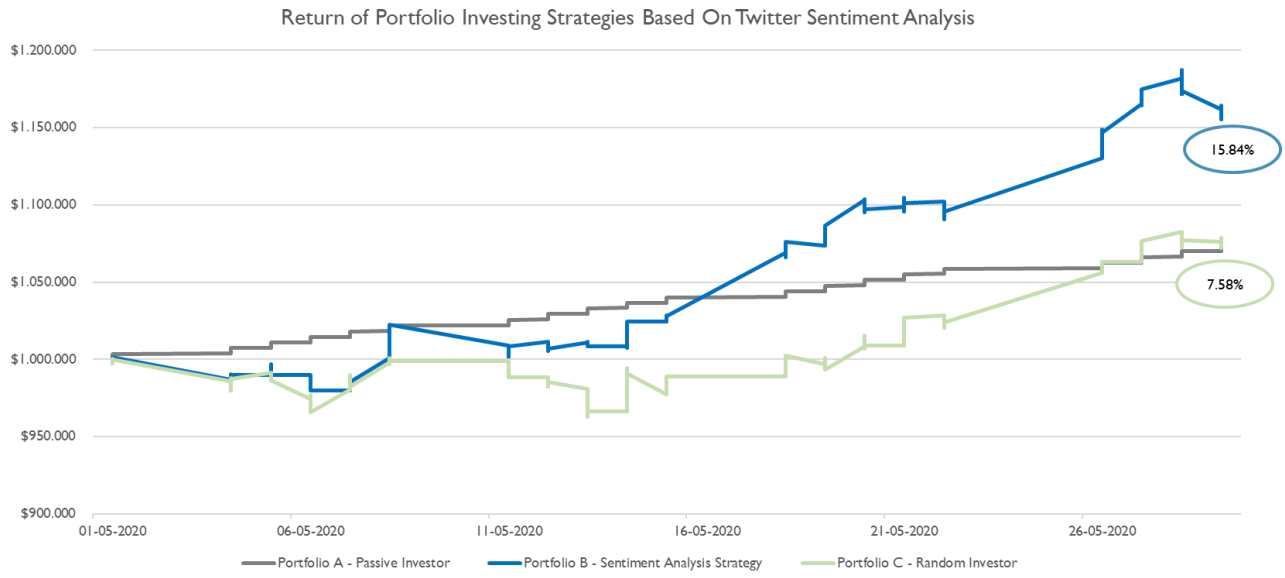


Figure 12: Graphical representation of the portfolios' performance during the testing period in May 2020. The performance of Portfolio B is noticeably better than the performance of the other two portfolios used as benchmarks.

This experiment conducted with the three portfolios aimed at verifying that the 58.25 % accuracy led to a meaningful improvement of the investment strategy. Indeed, if the model had correctly predicted only those 30 minutes periods during which the market movements were negligible, while wrongly predicting high absolute value market swings, then portfolio B would have gained a small amount of money during the correctly predicted periods, and lost important amounts during the wrongly predicted periods, despite the last ones being in minority. The fact that portfolio B's performance differs so much from the one of portfolio C suggests that the results obtained are significant.

Although this result shows that the model is able to correctly predict and lead to twice the return that one would normally have, if investing without the algorithm, there is one pivotal criticism to this method: it can not be used for business purposes, as it is. Indeed, the algorithm predicts an increase or a decrease in the security price by analysing all the tweets from the same time-period. In other words, the algorithm predicts the performance based on tweets posted from the beginning until the end of the 30 minutes period, hence would only be able to give a prediction after collecting all the data, therefore after the 30 minutes have passed. However, an investor would need to know the performance during those 30 minutes at the very beginning of the period, to be able to decide on whether to buy or sell the security. If the algorithm were able to really predict the performance of the security using the data collected in the 30 minutes timeframe before, then this would represent an extraordinary result for business applications. Thus, one last step in this research is represented by verifying what would happen if the algorithm were set to use tweets posted in the previous timeframe, to predict what would happen in the stock market in the next 30 minutes period.

Showing a dependency of the market performance during a 30 minutes period (period 2) from the period immediately before (period 1) is non-trivial. Indeed, this means that the market performance would still be linked to sentiments expressed up to 1 hour before: the security performance at the end of period 2 is predicted using data posted also at the beginning of period 1, hence 60 minutes before. In such a volatile context as the markets can be, 1 hour difference is significant. Therefore, it is not surprising to find out that SVM does not perform well in this case, with around 48 % accuracy. On the other hand, kNN performs well, although with a slight decrease in accuracy. As for the previous experiment, kNN with k=5 and the subset being "positive tweets/total tweets", "negative tweets/total tweets", "total tweets" and "positive tweets/negative tweets" performs the best with an **accuracy of 55.56 %** (vs. previous 58.25%) **on the XLI index.**

With such predictions, portfolio A's return is 7.33%, as nothing changed from before. Portfolio C's return is 8.23% using the random investment strategy. Portfolio B's return with the 30 minutes predictive strategy is 12.32 %. Therefore, this result suggests that tweets are indeed able to predict market movements with a 30 minutes time shift, confirming the relevance of the method applied.

It is worth noticing that this strategy is developed assuming that there are no transaction fees impacting the return of the portfolio, hence allowing for a purchase and sale of securities for as many times as required by the algorithm predictions. Moreover, the strategy is tested in a period of extremely high volatility, when markets were rebounding from low values attained during the COVID-19 pandemic outburst. Therefore, this is one of the reasons why the portfolio returns are so high even in only one month of strategy implementation. The strategy could be implemented in less volatile environments to observe the behaviour in different market conditions.

7. Future Development and Research

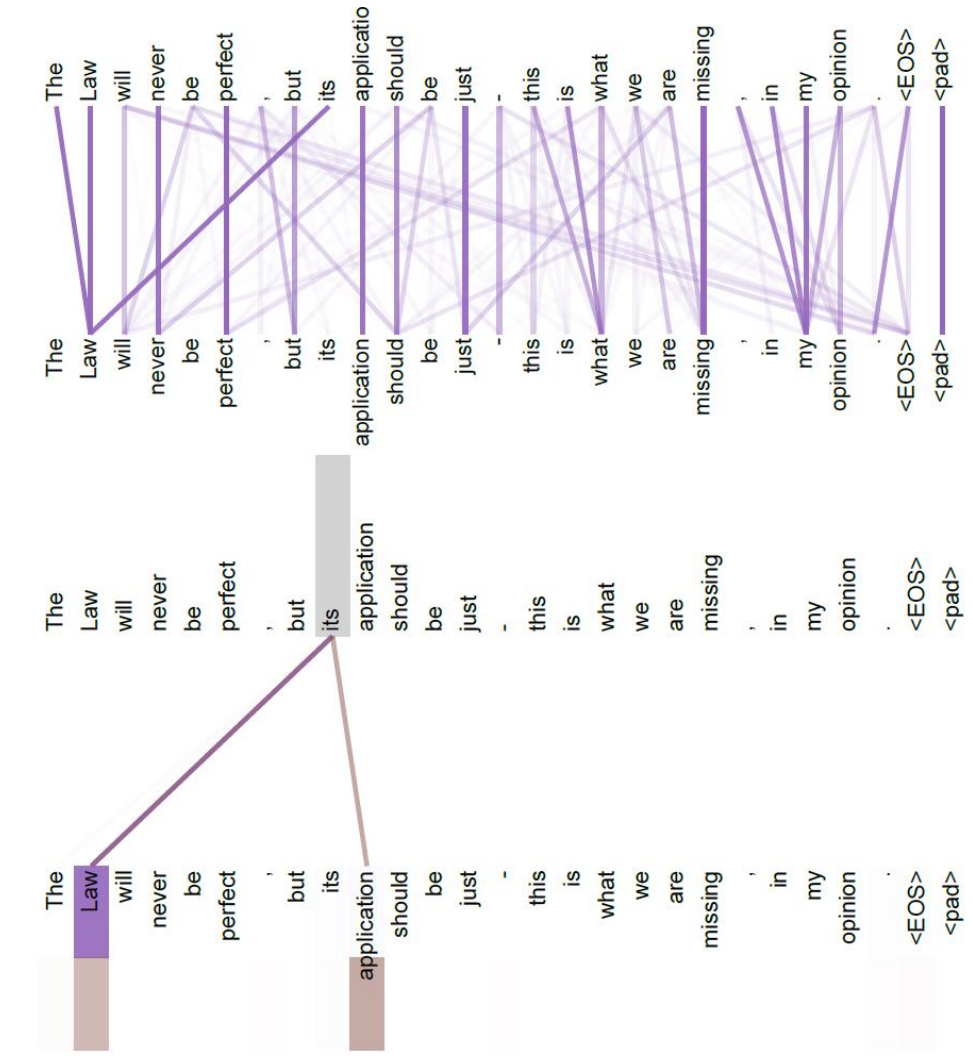
The results obtained in this research corroborate the previous findings discussed in the Literature Review. They suggest the possibility of improving a portfolio's return by using an investment strategy based on the sentiment analysis of tweets. Although these results constitute an important arrival point, however they can also be considered the beginning for analyses considering variations on different aspects. Further analysis could use a vaster dataset, exploring the correlation in other markets such as the UK, European countries and potentially extending the analysis period to include the second and third COVID-19 contagion waves. In particular, the second wave could be particularly interesting to study, since it had a strong impact on the population, who realized that the pandemic was not over and that it could last for a long period.

Another important improvement that could be made would be considering a finer granularity of the ETF dataset, in order to be able to observe closer effects of Twitter sentiment on the market. For example, 5-minutes or 1-minute intraday data are also available, for a higher cost, on some financial data websites. Also, different ETFs can be analysed, as well as considering different underlying securities such as single stocks, for example. Along with an improvement in the dataset, also an improvement of the model and algorithms could be useful. In particular, one could implement more complex algorithms to predict the increase or decrease of the security market price. This could be done by implementing random forest algorithms or neural networks structures with some layers of neurons. These models require powerful hardware, hence also hardware could be an aspect of improvement. Moreover, new models could differ from the ones applied in this research and they could allow for a prediction of the exact percentage increase or decrease in the security price, instead of predicting directional movements as in the research. This would allow advanced investment strategies to be put in place, also considering the transaction fees and only investing when the predicted return on the portfolio is higher than the fees paid for a transaction.

A further development idea would be to apply the same model and strategy on specific periods and contexts. For example, the same process can be applied to Elon Musk's tweets and the market performance of his company Tesla, of the company Twitter itself during the attempted acquisition or of the cryptocurrencies. Another extremely interesting topic would be analysing the Russian war against Ukraine as a topic on Twitter, linked to the overall markets' performance or to the energy commodities such as gas and oil prices.

The vastity of the research topic allows for many analyses with different variables. This is also one of the reasons why the topic of using social media sentiment analysis has been studied for a long time, but still attracts a lot of enthusiasm and energy. The results obtained in this research represent a positive outcome, which, if further explored, could find business applications.

8. Appendix



Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

Figure A1: example of Attention from the paper "Attention Is All You Need", Vaswani et al. (2017)

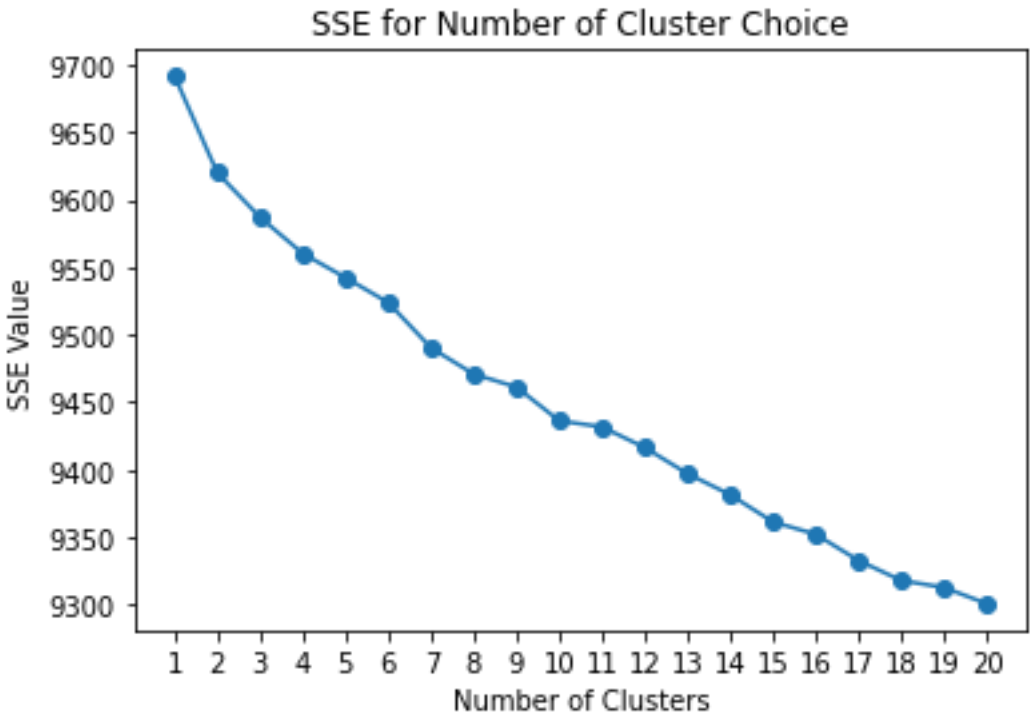


Figure A2: graph showing the SSE values for the different number of clusters used. It represents the sum of squared errors, hence the sum of squared distances between each element of the cluster and its centroid. It is used to determine the best number of clusters, which needs to be a trade off between the accuracy given by having many clusters and the overfitting and complexness of such algorithms.

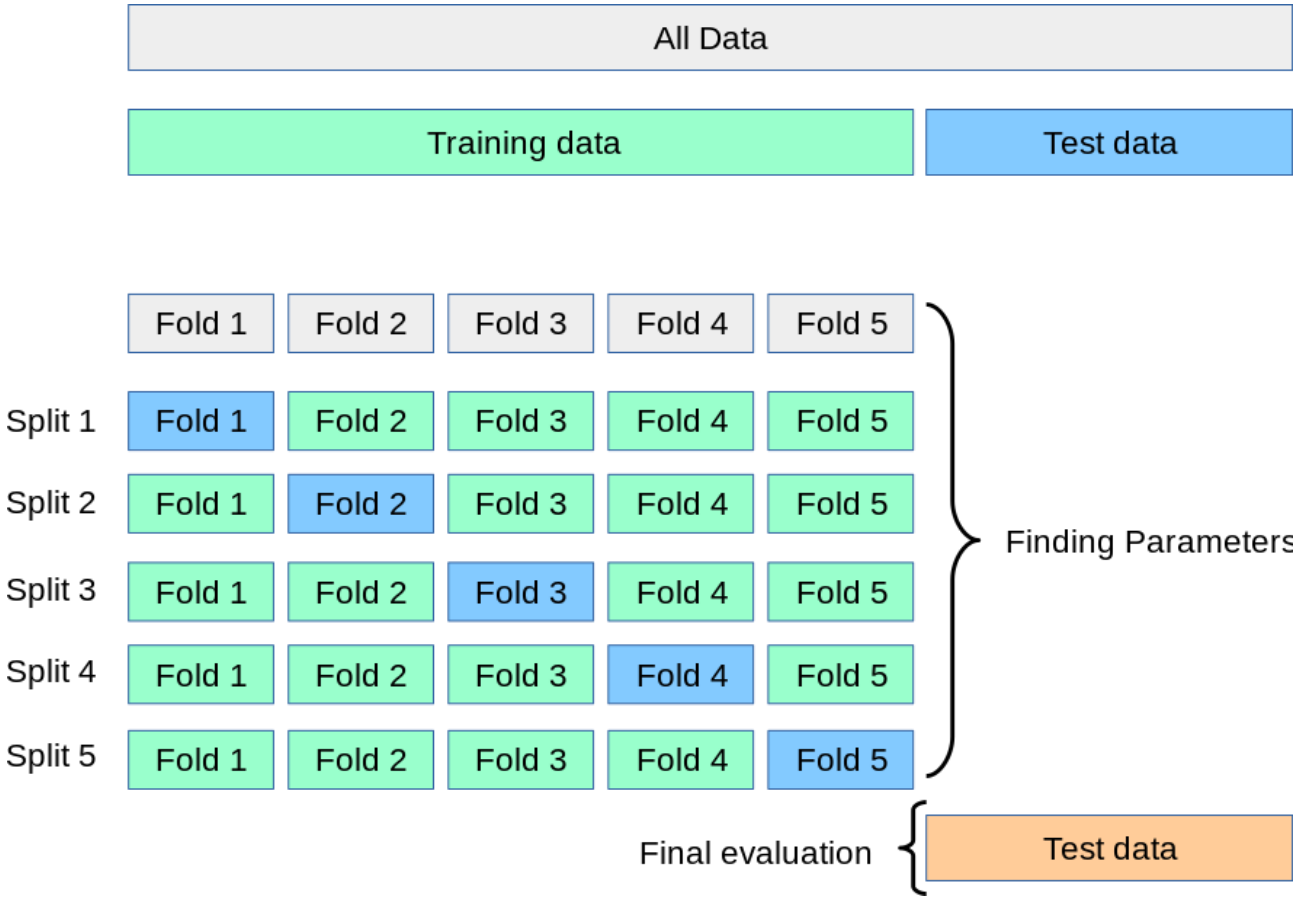


Figure A3: Example of 5-fold cross validation where the algorithm is finetuned on the whole dataset by performing 5 times the process of training and validating on 5 different subsets. Each "Fold" corresponds to 20% of the original dataset

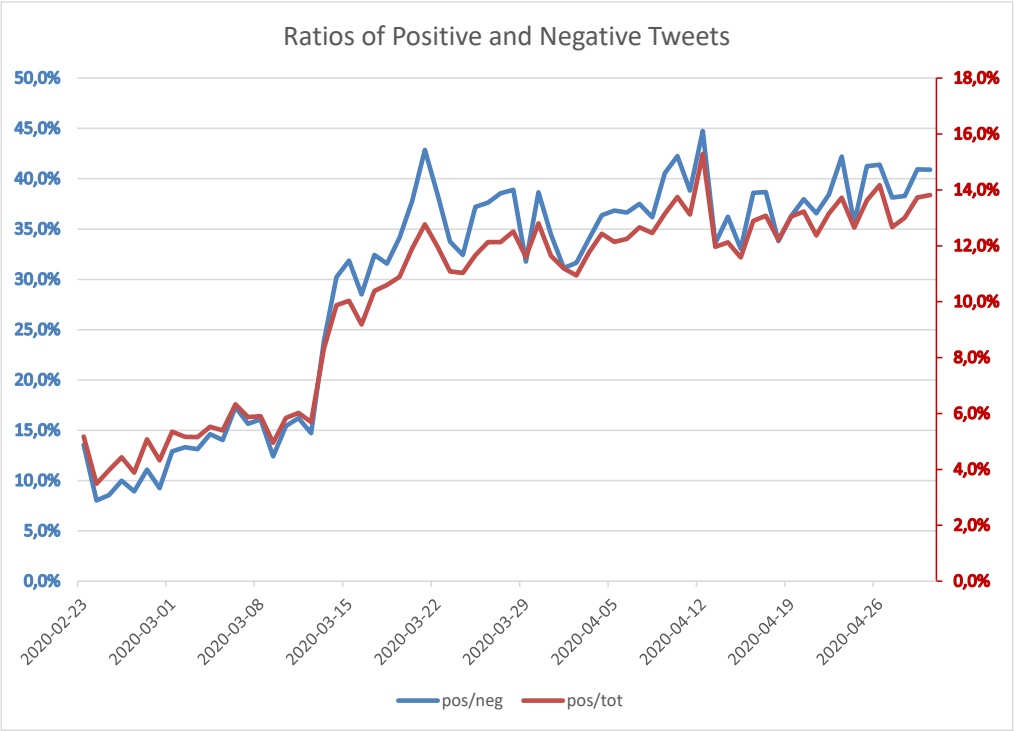


Figure A4: The graph represents the ratios between positive and negative sentiment tweets over the period of time from 23rd February until 30th April. It is important to notice how both the ratios positive/negative and positive/total were extremely low at the beginning of the period due to a high number of negative sentiment tweets. Over time the population started to understand what COVID-19 represented and expressed a slightly more positive sentiment.

Valuation and Test Results

Highlighted cells show the related feature combination leads to the highest result among all subsets with the same settings. E.g. for k=5 in kNN in XLI, 53,64% is the highest value reached among all features subsets

Highlights overall best result

ETF - XLI

Features	SVM Accuracy	kNN Accuracy			
		k = 1	*k = 3*	*k = 5*	*k = 7*
neg/tot, pos/neg, tot	53,50%	51,26%	51,26%	53,64%	53,09%
pos/neg, pos/tot, tot	53,50%	51,26%	51,26%	53,64%	53,09%
neg/tot, pos/neg, pos/tot, tot	53,50%	51,26%	51,26%	**53,64%**	53,09%

ETF - XLF

Features	SVM Accuracy	kNN Accuracy			
		k = 1	*k = 3*	*k = 5*	*k = 7*
neg/tot, pos/neg, surprise	51,88%	50,91%	52,43%	52,43%	51,88%
pos/neg, pos/tot, disgust	51,18%	49,93%	49,65%	52,16%	52,16%
pos/neg, pos/tot, fear	54,10%	52,29%	52,29%	52,29%	51,88%
neg/tot, pos/neg, pos/tot, surprise	51,88%	50,91%	52,43%	52,43%	51,88%
neg/tot, pos/neg, fear	54,10%	52,29%	52,29%	52,29%	51,88%
neg/tot, pos/neg, pos/tot, fear	54,10%	52,29%	52,29%	52,29%	51,88%

ETF - XLK

Features	SVM Accuracy	kNN Accuracy			
		k = 1	*k = 3*	*k = 5*	*k = 7*
neg/tot, pos/neg, fear	49,72%	48,96%	50,48%	51,18%	49,93%
neg/tot, pos/neg, pos/tot, fear	49,72%	48,96%	50,48%	51,18%	49,93%
neg/tot, pos/neg, anger	50,69%	47,56%	49,65%	50,21%	48,82%
neg/tot, pos/neg, sadness	52,50%	49,10%	47,57%	47,57%	47,15%

ETF - XLV

Features	SVM Accuracy	kNN Accuracy			
		k = 1	*k = 3*	*k = 5*	*k = 7*
neg/tot, pos/neg, sadness	53,41%	50,76%	50,34%	51,88%	51,88%
neg/tot, pos/tot, sadness	53,41%	50,76%	50,34%	51,88%	51,88%
neg/tot, pos/neg, fear	49,38%	49,10%	50,07%	51,88%	51,88%

ETF - XLI

Test	kNN	SVM
	k = 5	
neg/tot, pos/neg, pos/tot, tot	**58,25%**	52,19%

ETF - XLF

Test	kNN	SVM
	k = 5	
neg/tot, pos/neg, surprise	46,82%	55,52%
pos/neg, pos/tot, fear	55,52%	46,82%

ETF - XLK

Test	kNN	SVM
	k = 5	
neg/tot, pos/neg, sadness	47,32%	52,68%
neg/tot, pos/neg, anger	47,32%	57,38%

ETF - XLV

Test	kNN	SVM
	k = 5	
neg/tot, pos/neg, sadness	52,19%	47,14%
neg/tot, pos/tot, sadness	52,19%	47,14%
neg/tot, pos/neg, fear	52,19%	52,53%

Figure A5: SVM and kNN models results for the best performing subsets of features

9. Acknowledgements

I am grateful to Professor Tania Cerquitelli and Professor Alberta Di Giuli for their guidance and precious suggestions supervising my research. I also thank Professor Luca Cagliero for his advice and Doctor Paolo Bethaz for his help setting up the Twitter Research account.

Twitter data have been downloaded through a Twitter Developer Academic Research account (available at https://developer.twitter.com/en/products/twitter-api/academic-research). Financial information have been downloaded from Kibot (available at https://kibot.com). Computational resources are partially supported by SmartData@polito.

10. References

Al Guindy, M. (2021). Corporate Twitter use and cost of equity capital. Journal of Corporate Finance, p.101926. https://doi:10.1016/j.jcorpfin.2021.101926

Albuquerque, R., Koskinen, Y., Yang, S. and Zhang, C. (2020). Resiliency of Environmental and Social Stocks: An Analysis of the Exogenous COVID-19 Market Crash. *The Review of Corporate Finance Studies*, [online] 9(3), pp.593–621. https://doi:10.1093/rcfs/cfaa011

Baker, S.R., Bloom, N., Davis, S.J., Kost, K.J., Sammon, M.C. and Viratyosin, T. (2020). *The Unprecedented Stock Market Impact of COVID-19*. [online] National Bureau of Economic Research. Available at: https://www.nber.org/papers/w26945

Bloomberg. (20th February 2014) Trending on Twitter: Social Sentiment Analytics. Available at: https://www.bloomberg.com/company/press/trending-on-twitter-social-sentiment-analytics/ (Accessed: May 2022)

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp.1–8. https://doi:10.1016/j.jocs.2010.12.007

Campbell, R. H. (2020) Pandemic of 2020: Economic and Financial implications. Available at: https://faculty.fuqua.duke.edu/~charvey/Audio/COVID/COVID-Harvey.html (Accessed: May 2022)

Chen, H., De, P., Hu, Y. (Jeffrey) and Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, [online] 27(5), pp.1367–1403. https://doi:10.1093/rfs/hhu001

Davis E., Morgenstern L., Ortiz C. (n.d) The Winograd Scheme Challenge. Available at: https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html (Accessed: June 2022)

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805

Fahlenbrach R., Rageth K., Stulz R. M., How Valuable Is Financial Flexibility when Revenue Stops? Evidence from the COVID-19 Crisis, *The Review of Financial Studies*, Volume 34, Issue 11, November 2021, Pages 5474–5521, https://doi.org/10.1093/rfs/hhaa134

Gu, A. professor of finance C. and Kurov, P. of finance A. (2020). Informational Role of Social Media: Evidence from Twitter Sentiment. Journal of Banking & Finance, p.105969. https://doi:10.1016/j.jbankfin.2020.105969

Harford, J., Jiang, F., Xie, F., 2019. Analyst career concerns, effort allocation, and firms' information environment. Rev. Financ. Stud. 32 (6), 2179–2224.

Jay A. (2022) Number of Twitter Users 2022/2023: Demographics, Breakdowns & Predictions. Available at: https://financesonline.com/number-of-twitter-users/ (Accessed: June 2022)

Kilcher Y., (28th November 2017) YouTube: Attention Is All You Need. Available at: https://www.youtube.com/watch?v=iDulhoQ2pro&t=1333s (Accessed: May 2022)

Kibot (n.d.) Historical Intraday Data. Available at: https://kibot.com (Accessed: June 2022)

Knowledge Center, (n.d) BERT Encoder. Available at: https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/bert-encoder (Accessed: June 2022)

Kulshrestha R. (29th June 2020) Towards Data Science: Transformers. Available at: https://towardsdatascience.com/transformers-89034557de14 (Accessed: June 2022)

Mazur, M., Dang, M. and Vega, M. (2020). COVID-19 and the March 2020 Stock Market Crash. Evidence from S&P1500. *Finance Research Letters*, 38(101690), p.101690. https://doi:10.1016/j.frl.2020.101690

Mislove A., Lehmann S., Ahn Y., Onnela J., Rosenquist J. (2010), Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter. Available at: https://www.ccs.neu.edu/home/amislove/twittermood/ (Accessed: May 2022)

Mittal, A. and Goel, A. (n.d.). *Stock Prediction Using Twitter Sentiment Analysis*. [online] Available at: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

Narayan, P.K., Phan, D.H.B. and Liu, G. (2020). COVID-19 lockdowns, stimulus packages, travel bans, and stock returns. *Finance Research Letters*, 38, p.101732. https://doi:10.1016/j.frl.2020.101732

Nguyen, D., Vu, T. and Nguyen, A. (2020). *BERTweet: A pre-trained language model for English Tweets*. [online] pp.9–14. Available at: https://aclanthology.org/2020.emnlp-demos.2.pdf

Oliveira, N., Cortez, P. and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, pp.125–144. https://doi:10.1016/j.eswa.2016.12.036

Pérez J. M. (n.d.) Pysentimiento. Available at: https://huggingface.co/pysentimiento (Accessed: June 2022)

Perez, J.M., Giudici, J.C. and Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. [online] Available at: https://arxiv.org/pdf/2106.09462.pdf

Pi School, (4th October 2017) YouTube: Attention is all you need; Attentional Neural Network Models | Lukasz Kaiser | Masterclass. Available at: https://www.youtube.com/watch?v=rBCqOTEfxvg&t=1580s (Accessed: June 2022)

Pupale R. (16th June 2018) Towards Data Science: Support Vector Machines (SVM) – An Overview. Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989#:~:text=SVM%20or%20Support%20Vector%20Machine,separates%20the%20data%20into%20classes (Accessed: June 2022)

Scikit Learn (n.d.) Scikit Learn: k-Nearest Neighbours Classifier. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (Accessed: June 2022)

Scikit Learn (n.d.) Scikit Learn: Support Vector Machines. Available at: https://scikit-learn.org/stable/modules/svm.html (Accessed: June 2022)

SESAMm, Available at: https://www.sesamm.com/ (Accessed: May 2022)

Stevens P., Fitzgerald M., Imbert F. (12th March 2020) Stock market live Thursday: Dow tanks 2,300 in worst day since Black Monday, S&P 500 bear market. Available at: https://www.cnbc.com/2020/03/12/stock-market-today-live.html (Accessed: May 2022)

Surowiecki, J., 2004. The Wisdom of Crowds. Random House, New York

The Financial Times. GameStop and BlackBerry shares soar on amateur traders' fervour. Available at: https://www.ft.com/content/f8937ac7-da4a-41a6-bbbf-7b573e8ed72b (Accessed: May 2022)

Thomson Reuters. (3rd February 2014) Thomson Reuters Adds Unique Twitter and News Sentiment Analysis to Thomson Reuters Eikon. Available at: https://www.thomsonreuters.com/en/press-releases/2014/thomson-reuters-adds-unique-twitter-and-news-sentiment-analysis-to-thomson-reuters-eikon.html (Accessed: May 2022)

Twitter (n.d.) Academic Research access. Available at: https://developer.twitter.com/en/products/twitter-api/academic-research (Accessed: April 2022)

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] Available at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Yildirim S., (29th February 2020) Towards Data Science: k-Nearest Neighbours (kNN) – Explained. Detailed theoretical explanation and scikit-learn implementation. Available at: https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3 (Accessed: June 2022)