

Predicting stock prices through text media sentiment analysis: Memoria v1

Juan Luis Ruiz-Tagle

April 2020

Contents

1	Introduction and goals	3
1.1	Introduction	3
1.2	Related work	3
1.3	Problem statement and contribution	4
2	Theoretical background	5
2.1	Generative Adversarial Networks	5
2.2	Bidirectional Encoder Representations from Transformers	5
3	Development	5
3.1	Data gathering	5
3.1.1	Stock selection	5
3.1.2	News articles text data	6
3.1.3	Financial returns data	6
3.2	First iteration: Predicting returns based only on sentiment analysis	6
3.3	Second iteration: Using cGANs to predict returns	8
4	Results	8
5	Conclusions	8
6	Future work	8

1 Introduction and goals

1.1 Introduction

Stock market prediction is a very active research field due to the significant profits it can yield. With the advent of data science and deep learning combined with the increasing computing power capabilities we have at our disposal, stock market prediction has become more than ever a very promising research area. It is true that randomness is intrinsic to the very nature of stocks [1] but it is still feasible to find patterns and correlations which can partially predict stock trends in a general fashion.

Another field which has lately surpassed many milestones thanks to novel data science techniques is Natural Language Processing (NLP). Language understanding has evolved and become more perfect. Text translators perform better, bots seem more natural and accurate in their answers, speech recognition software captures a broader spectre of voices, etc.. TODO = poner papers para estos ejemplos.

A point of union between these two fields is the text from newspaper financial articles. This text expresses the current situation of the stock markets. In the other hand, it is main the source of information for many people which trade with stocks. There seems to be some sort of cyclical relationship between newspapers and stocks. If things go bad for a stock, newspapers will talk about it. If people read these bad news, fear will make them react in order to stop losing money or gain profit from the situation. The same reasoning follows if things go well, since good expectations will encourage people to buy. Of course, these premises do not hold true always. Money is not only invested by newspaper readers after the morning coffee. Many millions are invested daily following non-public business decisions, trading algorithms and other causes which escape the words of newspaper articles.

The aim behind this thesis is to reconcile stock market prediction with sentiment analysis of newspaper data, to see if we can predict future trends with help of sentiment analysis. First, related work on this topic is revised. Then, the problem statement together with the research question and the main contributions of the thesis will be defined. The next section covers theoretical reasoning about design choices. Then it comes the development of the architecture for the system, together with the tools used. Afterwards, research results are presented and conclusions are drawn. Finally, possible threads to continue the work beyond the thesis are proposed.

1.2 Related work

The invention of Generative Adversarial Networks (GANs) by Ian Goodfellow et. al. [2] in 2014 constituted a novel approach to solve many deep learning problems and rapidly became extremely popular. It merged ideas from deep learning and reinforcement learning, taking the best from each field. After some time, researchers started coming up with variants of the original GANs,

like the Wasserstein GAN (WGAN) [3], the Least Squares GAN (LSGAN) [4] or the Conditional GAN (cGAN) [5]. An overview of the different variations was published by Hong [6]. GANs are already being used in a broad range of fields. I am interested in exploring their potential to predict financial markets. Some studies have already been done in this topic, using GANs to estimate future returns, for example [7].

BERT (Bidirectional Encoder Representations from Transformers) is a neural net tasked to solve any kind of NLP problem. It was developed by researchers at Google [8] and it became the state of art, breaking records in many different NLP benchmarks. Language modelling networks are usually trained by randomly masking words in each sentence and trying to predict them given the previous or following ones. Given millions of ordered sentences, these networks learn to predict accurately the empty spaces looking at the text either before or after the masked word, but not both. If you train the network with information from both sides the model would overfit since you would be implicitly telling the answer to it when training with other sentences. This is where BERT's core novelty comes into play since it is designed to learn from the text before and after the masked words (that is what Bidirectional means). BERT overcomes this difficulty by using two techniques Masked LM (MLM) and Next Sentence Prediction (NSP)

TODO: explicar más en detalle algunos de los otros papers.

TODO: I cite everything here to make sure it appears in the bibliography (To be removed) [9] [10] [11] [12] [8] [2] [5] [7] [13] [14] [15] [16] [17] [6] [3] [4] [18] [1]

1.3 Problem statement and contribution

The research question I will be giving answer to with this thesis is: "Can we predict more accurately the returns of a stock with GAN based model if we include newspaper text data into it?"

The main contribution of the thesis is giving an answer to this question. The path which will take us there consists in taking the following steps:

- Investigate how to effectively predict market tendencies using exclusively sentiment analysis of newspaper text as input to the model. BERT will be used for this task. Due to the limitations of textual data, predictions won't be accurate but will still be significant
- Set up a GAN to predict the evolution of a stock for the next 10 days, using only historical financial data.
- Set up a cGAN with the same purpose, but adding as well the sentiment analysis of the textual data as conditional input.
- Compare the previous two models and check if there is an improvement in the model which includes textual data.

Volatility \ Trend	Increasing	Regular	Decreasing
High	Bitcoin (BTC USD)	Tesla (TSLA)	Kraft-Heinz (KHC)
Low	McDonalds (MCD)	PepsiCola (PEP)	Occidental Petroleum (OXY)

Table 1: Selected stock for our experiments

As a secondary contribution, we will design and test a novel loss function substituting the regular MSE, looking for more accurate predictions. To my best knowledge, these contributions are sufficiently genuine.

2 Theoretical background

2.1 Generative Adversarial Networks

TODO. Aquí pretendo hacer una descripción detallada del funcionamiento de las GANs, y en concreto de las conditional GANs.

2.2 Bidirectional Encoder Representations from Transformers

TODO. Descripción interna de BERT y del HuggingFace Transformer utilizado para extraer el sentimiento de las noticias.

3 Development

3.1 Data gathering

3.1.1 Stock selection

Five stocks were be selected to run our experiments, namely Tesla, Kraft-Heinz, PepsiCola, McDonalds and Occidental Petroleum. All of them belong to the S&P500. I was interested in focusing in big brands whose financial health would be covered in the newspapers. At the same time, they had to be representative of the whole index, that is, with different growing trends and volatility rates. Also I chose to add Bitcoin to the set of time series to be predicted. Its value has an enormous volatility and it is very prone to change by sudden hypes and fears, usually reflected in newspapers. Although Bitcoin is a cryptocurrency and not a stock, strictly speaking, it can be bought and sold in the same fashion. This makes it perfectly suitable for our needs. Table 1 depicts our stocks in relation to their growing trend and volatility rate.

The date interval from which we gather data (news articles and returns) goes from 1/1/2019 to the 19/3/2020. The models will be trained with data until the 1/11/2019 and tested from then on.

3.1.2 News articles text data

To retrieve articles related to the selected stocks I used two python packages which came very handy, namely `googlesearch` TODO: citar and `news-please` TODO: citar. The former emulates a search in google and retrieves a set of URLs, and the latter extracts a lot of information from an article (publication date, authors, main title, main text, etc....) given an URL.

Combining these two, I simulated searches in google news with the appropriate search terms for each stock. I did this for every day in the chosen date ranges and collected the top 5 articles about each stock written in English.

Being able to simulate a google search guarantees that the top articles that appear on the list are the most relevant ones, and one could guess that the ones that had more impact. Figure 1 showcases a small sample of the data gathered for Bitcoin.

	date_article	authors	title	description	maintext	source_domain
490	2019-04-09	Brenda Goh,Min Read	China wants to ban bitcoin mining	China's state planner wants to eliminate bitco...	SHANGHAI/HONG KONG (Reuters) - China's state p...	www.reuters.com
868	2019-06-24	Kate Rooney	Bitcoin rallies above \$11,000 through weekend,...	Bitcoin is approaching its highest level in mo...	Bitcoin is closing in on its highest level in ...	www.cnbc.com
351	2019-03-13	Samantha Chang	'Crypto' Movie Stirs Backlash by Pushing Bitco...	The new film 'Crypto' is a cyber-thriller that...	Hollywood already has a sinister opinion about...	www.ccn.com
355	2019-03-14	NaN	Bitcoin is Cheap Until April, We'll Never See ...	Despite being down 80 percent from its all-tim...	Despite being down 80 percent from its all-tim...	www.ccn.com
1563	2019-11-10	William Suberg	Bobby Lee: \$500K Bitcoin Price 'Flipping' of...	A price of \$500,000 is easily in reach within ...	Bitcoin (BTC) will surpass the market cap of g...	cointelegraph.com
1914	2020-01-19	Christina Comben,Trevor Smith	Sunday Digest: Bitcoin Price, BSV Pump and Dum...	If last week, bitcoin price was all about \$8k...	Bitcoin Sunday Digest: Bitcoin Price, BSV Pump...	bitcoinist.com
1746	2019-12-17	Liam Stack	Unable to Retrieve Money, Cryptocurrency Inves...	Gerald W. Cotten, the C.E.O. of Quadriga CX, w...	There's Bitcoin. There's Litecoin. There's Eth...	www.nytimes.com
1683	2019-12-04	Matthew Beedham,December	Drugs hidden in child's toy lead police to mas...	NaN	Australian police have reportedly seized a rec...	thenextweb.com
1051	2019-07-31	Kevin Helms,A Student Of Austrian Economics,Ke...	Indian Finance Minister Addresses Crypto Proposal	India's finance minister has broken silence an...	Indian Finance Minister Addresses Crypto Propo...	news.bitcoin.com
1276	2019-09-14	Christina Comben,Trevor Smith	German Gov't Approves 'Bundes-Chain' to Combat...	The German government will approve its propose...	News teaser German Gov't Approves 'Bundes-Chai...	bitcoinist.com

Figure 1: Sample of Bitcoin news dataset

3.1.3 Financial returns data

To obtain returns for the selected stocks, I downloaded the close price values using the python package `yfinance`. TODO: citar

3.2 First iteration: Predicting returns based only on sentiment analysis

After collecting the news articles data, all the pieces of text were split into sentences so, in the end, my dataset consisted of a bunch of sentences grouped by the day they were published. Then I used a version of BERT available as a [HuggingFace transformer](#) which is pretrained to do sentiment analysis on product reviews.

Given a product review, it predicts its "sentiment" as a vector of number of stars (1 to 5). Reviews which are negative will have very high values for

the indices representing low stars and low values on the other ones. On the contrary, positive reviews will have high values for high stars and low values for the others. Even though product review text and newspaper text are fairly different, we will see that this model works surprisingly well on newspaper data. For example, if we input the sentence "Bitcoin futures are trading below the cryptocurrency's spot price" to the BERT HuggingFace Transformer, it returns the vector $[0.621, 0.741, 0.599, -0.509, -1.212]$. This prediction means that it is a 2-star text (the second value is the highest of the vector), that is, the text is slightly negative. Note that we are keeping all the values and not just classifying each sentence with a value from 1 to 5.

TODO: Desarrollar este parte con detalle y explicar mejor.

Resumen:

- Dividimos los artículos en frases, agrupandolas por día de publicación. Suponemos que tenemos N frases agrupados en D días.
- Evaluamos todas las frases obteniendo un vector de 5 estrellas para cada frase. Tenemos una matriz S de valoraciones con dimensiones $(N, 5)$
- Agrupamos las valoraciones por día, y calculamos la probabilidad media de cada estrella, obteniendo una matriz final de dimensiones $(D, 5)$
- Entrenamos una red LSTM con una ventana de 10 días y las 5 estrellas como features. Los labels son los retornos del activo para el día siguiente. Input shape $(N - 10, 10, 5)$ Output shape (N)
- Obtenemos los siguientes resultados:

TODO: mover estos gráficos a la sección de Results. Hacer comentario Precio del Bitcoin construido a partir de los retornos predecidos.

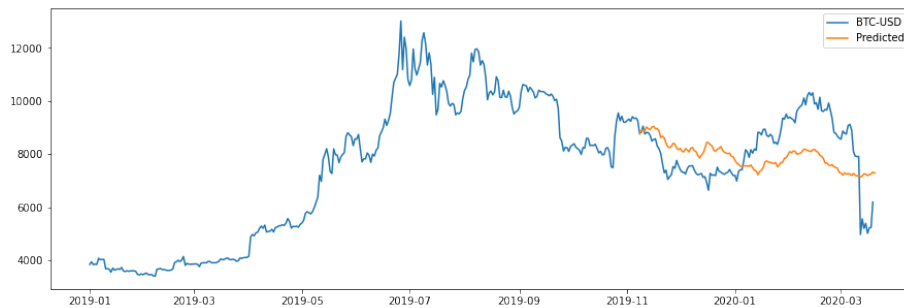


Figure 2: Bitcoin predicted price based solely on text sentiment analysis

Precio de la acción de Tesla construido a partir de los retornos predecidos.
El código para el desarrollo de esta parte se puede ver en mi [Github](#)

3.3 Second iteration: Using cGANs to predict returns

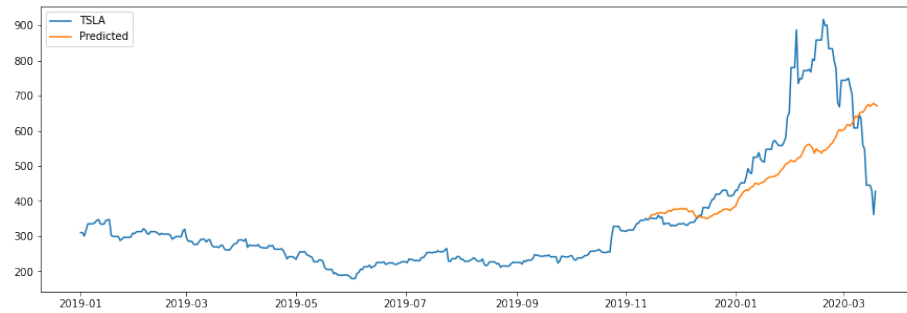


Figure 3: Tesla returns prediction based solely on text sentiment analysis

3.3 Second iteration: Using cGANs to predict returns

TODO: Explicar el ensamblaje de las cGANs en las que se usan retornos pasados y se ayuda del sentiment analysis para hacer predicciones de los retornos.

4 Results

5 Conclusions

6 Future work

References

- [1] A.-H. Sato and H. Takayasu, “Dynamic numerical models of stock market price: from microscopic determinism to macroscopic randomness,” *Physica A: Statistical Mechanics and its Applications*, vol. 250, no. 1-4, pp. 231–252, 1998.
- [2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.
- [3] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” 2017.
- [4] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” 2016.
- [5] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014.
- [6] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, “How generative adversarial networks and their variants work,” *ACM Computing Surveys*, vol. 52, p. 1–43, Feb 2019.
- [7] X. Zhou, Z. Pan, G. Hu, S. Tang, and C. Zhao, “Stock market prediction on high-frequency data using generative adversarial nets,” 2018.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.
- [9] T. Nguyen and K. Shirai, “Topic modeling based sentiment analysis on social media for stock market prediction,” vol. 1, 07 2015.
- [10] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis : A survey,” 2018.
- [11] C. Dos Santos and M. Gatti de Bayser, “Deep convolutional neural networks for sentiment analysis of short texts,” 08 2014.
- [12] S. Xiao, M. Farajtabar, X. Ye, J. Yan, L. Song, and H. Zha, “Wasserstein learning of deep generative point process models,” 2017.
- [13] T. Li, “Semi-supervised text regression with conditional generative adversarial networks,” 2018.
- [14] N. Dereli and M. Saraclar, “Convolutional neural networks for financial text regression,” 07 2019.
- [15] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” 2014.
- [16] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” 2019.

REFERENCES

- [17] N. Houlsby, A. Giurciu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” 2019.
- [18] J. Kalyani, P. H. N. Bharathi, and P. R. Jyothi, “Stock trend prediction using news sentiment analysis,” 2016.