# Fine-tuned AlBERTo for Stock Price Prediction: a Gibbs Sampling Approach

F. Colasanto[1], L. Grilli[2], D. Santoro[3] and G. Villani[3]

Big Data and Machine Learning in Finance Conference Politenico di Milano

June 10-11, 2021

[1] Scuola Internazionale Superiore di Studi Avanzati SISSA - Trieste, Italy
[2] University of Foggia, Dept. of Economics, Management and Territory - Foggia, Italy
[3] University of Bari Aldo Moro, Dept. of Economics and Finance - Bari, Italy

1. Introduction
2. Sentiment score
3. Markov Chain Monte Carlo
4. Conclusions

# Objective

The aim of the research is to improve the forecast in a very short time through polarity scores on the information present in the markets.

1 Introduction

2 Sentiment score
   BERT
   AlBERTo

3 Markov Chain Monte Carlo

4 Conclusions

1 Introduction

2 Sentiment score
- BERT
- AlBERTo

3 Markov Chain Monte Carlo

4 Conclusions

# Word Embedding

Word Embedding systems (such as word2vec and GloVe [*Pennington et al., 2014*]) that represent the basis of Natural Language Processing (NLP) are pre-trained on a certain corpus based on certain statistical relationships between occurrences.

# Word Embedding

Word Embedding systems (such as word2vec and GloVe [*Pennington et al., 2014*]) that represent the basis of Natural Language Processing (NLP) are pre-trained on a certain corpus based on certain statistical relationships between occurrences. However, these systems do not take into account "what is intrinsic in the text" (non-explicit relationships between words, what is implied).

# Architecture pre-BERT

Some of the most used architectures, such as ELMo or ULMfit are based on the use of biLSTM architectures. Accuracy (F1) is significantly improved over Word Embedding systems, but these systems still have problems related to separate training.

# Transformer

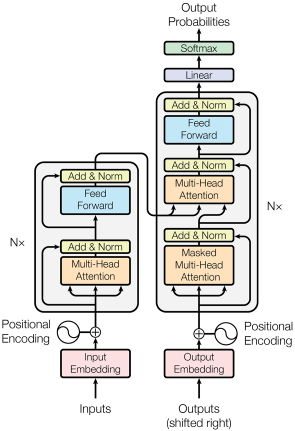



Figure: Transformer architecture [*Vaswani et al., 2017*]

- Multi-headed self attention;
- Feed-forward layers;
- Layer norm and residuals;
- Positional embeddings.

# BERT

Bidirectional Encoder Representations from Transformers (BERT) solve the problem of bidirectional language understanding.
The goal was to have a truly bidirectional lossless flow of information in a deep model.

# BERT

Bidirectional Encoder Representations from Transformers (BERT) solve the problem of bidirectional language understanding.
The goal was to have a truly bidirectional lossless flow of information in a deep model.

The solutions to this problem are to mask 15% of the words and make the prediction [Masked LM] and learn relationship between sentences, where B is the sentence that proceeds A or a random sentence [Next Sentence Prediction].

1. Introduction
2. Sentiment score
   - BERT
   - AlBERTo
3. Markov Chain Monte Carlo
4. Conclusions

# AlBERTo

AlBERTo is a BERT language understanding model for the Italian language *[Polignano et al., 2019]*. It is a system focused on masked learning, and is based on a large corpus of Italian Tweets *[Basile et al., 2018]* from 2012 to 2018 (BERT_base version).

https://github.com/marcopoli/AlBERTo-it

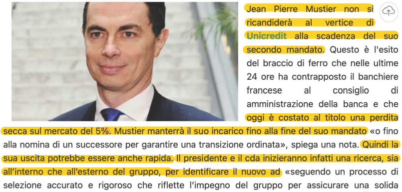
# Fine-tuning AlBERTo

Since AlBERTo is trained on a corpus of Tweets, we have decided to optimize it on the economic/financial world. This was possible by creating an additional layer on AlBERTo, training him through the *FinancialPhraseBank [Malo et al., 2014]* dataset (properly translated). This allowed us to maintain the accuracy of the previous model but at the same time refine its capabilities.

## Stocks

- **Unicredit S.p.A.**, in an article of 29/11/2020. The main topic concerns the failure to reapply the company's CEO, Jean Pierre Mustier, to his current position;
- **EssilorLuxottica S.A.**, in an article of 27/11/2020. The main topic concerns the possible acquisition of a Canadian company in order to make Luxottica's presence consolidated;
- **Intesa SanPaolo**, in an article of 30/7/2020. The main topic concerns an analysis of the bank's performance by Mediobanca analyzes and forecasts for the future following its merger with Ubi;
- **UnipolSai Assicurazioni**, in an article of 13/11/2020. The main topic concerns the analysis of the management of the insurance company in the quarters Q1-Q3 of 2020.

# Stocks



Jean Pierre Mustier non si ricandiderà al vertice di Unicredit alla scadenza del suo secondo mandato. Questo è l'esito del braccio di ferro che nelle ultime 24 ore ha contrapposto il banchiere francese al consiglio di amministrazione della banca e che oggi è costato al titolo una perdita secca sul mercato del 5%. Mustier manterrà il suo incarico fino alla fine del suo mandato «o fino alla nomina di un successore per garantire una transizione ordinata», spiega una nota. Quindi la sua uscita potrebbe essere anche rapida. Il presidente e il cda inizieranno infatti una ricerca, sia all'interno che all'esterno del gruppo, per identificare il nuovo ad «seguendo un processo di selezione accurato e rigoroso che riflette l'impegno del gruppo per assicurare una solida

(a) Extract of article (Unicredit)

EssilorLuxottica non si ferma al dossier GrandVision. E come anticipato oggi da *MF-Milano Finanza* ***il colosso mondiale dell'occhialeria sta per definire*** una nuova acquisizione internazionale per rafforzare la sua rete di store a livello globale.

Il target individuato, secondo quanto si apprende in ambienti finanziari, è un gruppo canadese quotato alla borsa di Toronto che conta su quasi 400 punti vendita sul mercato locale. L'identikit è quello di New Look Vision, azienda di Montréal fondata nel 1986 da Guy Rouleau che attualmente ha una capitalizzazione di 477 milioni di dollari canadesi e che è uno dei leader nazionali nel settore con 390 negozi. L'azionista di riferimento, con il 30%, è Benvent Holding.

(b) Extract of article (EssilorLuxottica)

# Sentiment score

| Stocks | Polarity | Avg. score |
|---|---|---|
| Unicredit | Negative | -0.50 |
| EssilorLuxottica | Positive | 0.32 |
| Intesa SanPaolo | Positive | 0.19 |
| UnipolSai | Positive | 0.25 |

Table: Stocks with sentiment prediction

1 Introduction

2 Sentiment score

3 Markov Chain Monte Carlo

4 Conclusions

# Geometric Brownian Motion

We can start from the model for the stock price dynamics:

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \quad (1)$$

Considering $Y = (S_1, ..., S_T)$ and $p(\mu, \sigma^2 | Y)$ the posterior distribution, we know that *[Johannes and Polson, 2003]*

$$\begin{aligned} p(\mu|\sigma^2, Y) &\propto p(Y|\mu, \sigma^2)p(\mu) \\ p(\sigma^2|\mu, Y) &\propto p(Y|\mu, \sigma^2)p(\sigma^2). \end{aligned} \quad (2)$$

# Markov Chain Monte Carlo

Assuming prior $p(\mu) \sim \mathcal{N}$ and $p(\sigma^2) \sim \mathcal{IG}$, the posterior are conjugate so $p(\mu|\sigma^2, Y)$ and $p(\sigma^2|\mu, Y)$ have also Normal and Gamma Inverse distribution respectively. In this case, we can use a Gibbs Sampler to produce a sample from the posterior distribution.

# Stock price

Since market information has different timing of impact on stocks, we consider the hourly price in order to have a forecast in 45 hours later (about 5 trading days after).

# Stock price

Since market information has different timing of impact on stocks, we consider the hourly price in order to have a forecast in 45 hours later (about 5 trading days after).
We can assume that the mean of daily returns is of order $10^{-4}$; while the volatility is of order $10^{-2}$, so that we have prior expectation for $\sigma^2$ equal to $0.01$.

# Stock price

Since market information has different timing of impact on stocks, we consider the hourly price in order to have a forecast in 45 hours later (about 5 trading days after).
We can assume that the mean of daily returns is of order $10^{-4}$; while the volatility is of order $10^{-2}$, so that we have prior expectation for $\sigma^2$ equal to $0.01$.

| N | $m$ | $\sigma_m$ | $\alpha_0$ | $\beta_0$ |
|---|---|---|---|---|
| 12000 | 0 | 0.0001 | 2.01 | 0.01 |

# Bounded parameters

- *Unicredit* stock: normal distribution support for mean bounded to $[-0.00060, -0.00040]$, while inverse gamma distribution support for volatility bounded to $[0.02, 0.06]$;
- *EssilorLuxottica* stock: normal distribution support for mean bounded to $[0.00022, 0.00042]$, while inverse gamma distribution support for volatility bounded to $[0.01, 0.05]$;
- *Intesa SanPaolo* stock: normal distribution support for mean bounded to $[0.00009, 0.00029]$, while inverse gamma distribution support for volatility bounded to $[0, 0.03]$;
- *UnipolSai* stock: normal distribution support for mean bounded to $[0.00015, 0.00035]$, while inverse gamma distribution support for volatility bounded to $[0.01, 0.05]$.

# Gibbs Sampling

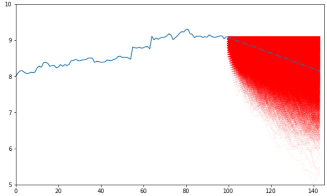
| | Unicredit | EssilorLuxottica | Intesa SanPaolo | UnipolSai |
|---|---|---|---|---|
| Without Gibbs Sampling | | | | |
| Mean | -0.000177 | -0.000066 | -0.00007 | -0.000086 |
| Standard dev. | 0.011029 | 0.010021 | 0.008952 | 0.007546 |
| Drift | -0.005244 | -0.000712 | -0.001356 | -0.002602 |
| Volatility | 0.073994 | 0.067222 | 0.060053 | 0.050622 |
| **With Gibbs Sampling** | | | | |
| Mean ($\hat{r} = 1$) | 0.000451 | 0.000242 | 0.000109 | 0.000177 |
| St. dev. ($\hat{r} = 1$) | 0.020014 | 0.010147 | 0.008953 | 0.010012 |
| **Drift** | **-0.02049** | **0.01092** | **0.00496** | **0.00803** |
| **Volatility** | **0.1343** | **0.0681** | **0.0601** | **0.0672** |

Table: Comparison between drift and volatility values before and after Gibbs sampling

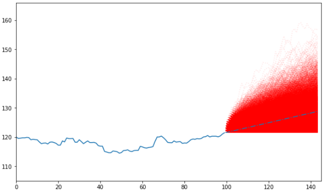
## Monte Carlo Simulations

Since we are interested in predicting stocks price, we have considered the following mechanism: using the bounded values of drift and volatility of the previous table we have generated a series of paths ($10000$) through Monte Carlo simulation, selecting the paths based on this polarity: if the news is positive we consider those whose final price ($S_T$) is included in the $[S_0, +\infty[$ range, otherwise, if the news is negative, we consider those whose final price is in the $]-\infty, S_0]$ range. The predicted final price is the expectation of the final prices in one of the two ranges (based on polarity).

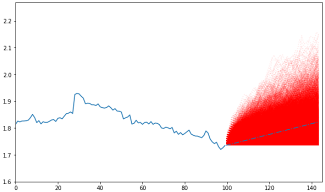
# Monte Carlo Simulations



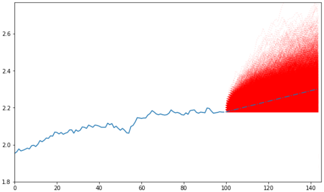

(a) Unicredit




(b) EssilorLuxottica




(c) Intesa SanPaolo




(d) UnipolSai

## Comparison

| Stocks | $S_0$ | $S_T$ | Real price | Variation |
|---|---|---|---|---|
| Unicredit | 9.094 | 8.14 | 8.10 | 0.0049 |
| EssilorLuxottica | 121.45 | 128.75 | 126.30 | 0.0193 |
| Intesa Sanpaolo | 1.735 | 1.823 | 1.8138 | 0.0050 |
| UnipolSai | 2.176 | 2.3017 | 2.2920 | 0.0042 |

Table: Comparison between predicted and actual prices (Variation $= \frac{|S_T - Real\ price|}{Real\ price}$)

1 Introduction

2 Sentiment score

3 Markov Chain Monte Carlo

4 Conclusions

## Future work

- Our main idea is not to use the punctual price, but to take advantage of the predicted price 5 days in advance;
- Use the predicted price in this way to improve portfolio optimization models (e.g. in Black and Litterman model as a fine-tuned views).

[Basile et al., 2018] V. Basile, M. Lai, and M. Sanguinetti.
Long-term social media data collection at the university of turin.
*Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6, 2018.

[Johannes and P., 2003] M. Johannes and N. Polson.
Mcmc methods for continuous-time financial econometrics.
2003.

[Malo et al., 2014] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala.
Good debt or bad debt: Detecting semantic orientations in economics texts.
*Journal of the Association for Information Science and Technology*, pages 782–796, 2014.

[Polignano et al., 2018] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile.
Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets.
*Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, 2019.

[Pennington et al., 2014] J. Pennington, R. Socher, and C. D. Manning.
Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[Vaswani et al., 2014] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin.
Attention is all you need.
*arXiv:1706.03762*, 2017.

# Thanks for the attention!

*for more details, please contact me:*

*domenico.santoro@uniba.it*