Financial Forecasting using Natural Language Processing Methods: A Literature Review

ELEC-E5550 - Statistical Natural Language Processing

Bruce Nguyen
School of Science
Aalto University
Espoo, Finland
bruce.nguyen@aalto.fi

Laurine Burgard-Lotz
School of Electrical Engineering
Aalto University
Espoo, Finland
laurine.burgard-lotz@aalto.fi

***Abstract*—Financial forecasting, or in a more applied sense, market prediction, has long been an attractive research problem. In fact, it is a very challenging topic due to the highly stochastic nature of markets, as well as the sheer number of response variables. Recently, the explosion of textual data and major advancements in natural language processing (NLP) has presented new promising approaches. As such, the use of NLP has become increasingly popular in this task, establishing the field of natural language based financial forecasting (NLFF). The techniques employed within NLFF largely fall within 2 categories: *event embeddings* and *sentiment analysis*. The use of the latter ones are based on the relationships between words and positive or negative opinions, opinions which naturally have an impact on the price of the stock market.**

In this review, we aim to give a brief context of the problem and then examine the latest developments in the field. The content will be structured following the categories of the most prevalent NLFF methods employed.

I. Introduction

The Efficient Market Hypothesis (EMH) stated that assets price reflect all the available information [9]. It follows that no one can expect to forecast the market with more than 50% accuracy. Even though this theory has been widely influential, both theoretical and empirical research has called in to question its validity [4]. Correspondingly, a body of stock price predicting techniques have grown considerably over time, forming the two major schools of thought in investing: technical and fundamental analysis [1].

Technical analysts attempt to predict market trends using its past behaviours, employing techniques from chart to cycle analysis. Fundamental analysis, on the other hand, is about measuring asset value by examining related economic and financial factors. Both practices are used in combination by investors, with the trend going towards technical analysis since the 1980s [16]. With regards to sources of information, fundamental analysts often rely on textual data to determine the "intrinsic" value of the target stocks. Thus, there has long been a desire to automatically analyze such documents using powerful computational methods in a timely manner, securing a competitive edge in the trading world.

Accordingly, attempts to achieve this objective have been recorded since as early as the 1980s, with the main focus on deriving semantic information at the time. However, most of them only use basic statistical NLP techniques such as word and word-coocurrence counting [10]. And, the most popular language model at the time was the primitive bag-of-words which lacks deep understanding. As such, the accuracy level was not sufficient for any application, and the scope was narrowly bounded around structured textual data [21].

From the last decade on-wards, developments across many domains have come together, presenting new avenues for NLFF research. First, the internet has brought about an explosion of real-time textual data for training machine learning models: for example, Twitter, or online forum discussions. Then, major leaps in deep learning methods helped NLP techniques to extract much more accurate semantic and sentiment knowledge. In fact, NLP systems are approaching human-level language understanding [19]. This lead to a surge in the number of papers looking to utilize these new advancements to model financial markets.

Before we proceed further in the subject, it is important to clarify that, even though financial forecasting encompasses ideas from credit scoring to inflation rate prediction [17], the majority of NLFF studies are dedicated to predicting the stock market and foreign exchange rate (FOREX). Xing et al. gave three reasons for this practice: (i) other assets' lack of accessibility, (ii) unique complexity and volatility of stock prices and FOREX, and (iii) their transparency [21]. Therefore, we take liberty to use the term market prediction interchangeably with financial forecasting.

Overall, there are two most popular types of approaches currently: event embeddings and sentiment analysis. The former mainly involves transforming financial events to vector representation, commonly performed on financial news and documents. Its rationale is simply news events affect human decisions and subsequently stock prices, as shown by Ding et al. [7]. This approach can be considered as the advanced version of semantic analysis using bag-of-words models that we have seen, adding structures to the otherwise ambiguous features.

For example, (Actor = Epic Games, Action = sues, Object = Apple) is a far more informative representation compared to {"Epic", "Games", "sues", "Apple."}. Regarding sentiment analysis, we know that emotions shaped human behaviours in addition to information. In fact, it has been shown that moods and sentiments significantly affect our financial decisions [15]. As such, analyzing public sentiment from news and social media to predict the market has yielded significant results empirically.

II. Methods

A. Event embeddings

Any discussion about event embeddings starts with the domain of knowledge graph embedding (KGE). In this area, researchers aim to represent and encode entities and relations between them as d-dimensional vectors and matrices, respectively - very much similar to how event embeddings work. Before the technique is appropriated into NLFF, KGE learning is most often done by minimizing a global loss function over the entities and relations in a knowledge graph [12, 2]. Most recently, techniques which use deep neural networks and low-dimensional hyperbolic space were explored [3].

Moving on to event embeddings in NLFF, its first use can be found in Ding et al. [8] as far as we know. Specifically, the authors used Open Information Extraction techniques to extract structured events from more than 10 million Reuters and Bloomberg financial news articles. This is then used by linear and nonlinear models such as Support Vector Machines and Neural Networks to examine events' influence on the stock market. In the end, the results gathered from the studies are: (i) structured events, or event embeddings, outperformed bags-of-words, (ii) deep neural network models are effective, (iii) the approach provided stable results, and (iv) data quality is important. Novel as method is in NLFF, the individual techniques employed remain rudimentary with respect to the wider context - notable is the use of a simple vanilla neural network. And, this is then addressed by Ding et al. [6] with a convolutional neural network (CNN) which takes the input of event embeddings. The model is then used to model the influence of both long-term and short-term events on the stock market, achieving an accuracy of up to 65%. Vargas, de Lima, and Evsukoff [20] further expanded on this by combining a recurrent layer to a CNN and thus creating a recurrent convolutional neural network (RCNN). In addition, the author also utilized word2vec - which is known as a major improvement in word embedding technique - in the process of generating event embeddings.

Later works shifted the focus to combining event embeddings with other useful features for the task, while also taking into account the interdependency. Hu et al. [11] drew on the human learning process to create a sophisticated model which takes into account (i) sequential context dependency, (ii) quality of information and (iii) effective and efficient learning. This was achieved by using a hybrid attention network that uses attention mechanisms at both temporal level and news level. After running simulations on various subsets of methods, the paper showed the strategy can generate an annualized return of 0.6. In another line of work, Zhang et al. [22] used not only news information but also public sentiment on social media. The authors also explored coupled matrix and tensor factorization scheme to support information integration, while taking into account correlations among features. Regarding event embeddings specifically, they are created using k-means methods from word embeddings - a novel approach - with the Chinese finance news corpus. The result is an accuracy of roughly 60%.

In recent years, Natural Language Processing models have continued to improve, allowing words to be represented more reliably. Many different tools for sentiment analysis already exist : **lexicon based methods** such as NLTK tool with VADER or TextBlob, **deep learning methods** with Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Gated Recurrent Units network (GRU), or even **Bidirectional Encoder Representations from Transformers** (BERT), ... These tools allow us to determine the opinion of an author towards a subject.

The field of finance is no exception: in the 2010s, a common method for sentiment analysis was based on lexicon analysis. A *sentiment lexicon* is a list of words which are labelled with their sentiment orientation, either positive or negative. In 2011, Loughran and McDonald introduced a really popular special lexicon for finance [13]. Lexicon-based methods then consists in calculating orientation for a document (more positive or more negative) from the semantic orientation of lexical features in the document, by using dictionaries of labelled words. They are expensive because they require a lot of work to build manually the lexicon, while statistical methods are easier to implement but generally give poorer results depending on the quantity of training data. This is why these lexicon methods are still used in Finance today, as shown in a recent paper by Sohangir, Petty, and Wang, who demonstrated that VADER (Valence Aware Dictionary for Sentiment Reasoning) outperform machine learning methods in extracting opinion from financial textual data (94.4% accuracy against 80.8% for the naive Bayes Classifier in the Stocktwits dataset), but is also faster.

Nonetheless, transformer models have established themselves as state of the art models for Natural Language Processing tasks. The so-called BERT model, for "Bidirectional Encoder Representations from Transformers", is a publicly available pre-trained Transformer model developed by Google in 2018 [5], which is able to outperform human scores in several NLP tasks. Nemes and Kiss used this new generation word embedding model in 2021 to predict stock values changes on the basis of stock news headline [14]. Their paper shows the superiority of BERT model by comparing it to TextBlob, NLTK-VADER Lexicon and Recurrent Neural Network (RNN).

What we want to address in our project is, how to use BERT for doing sentiment classification tasks in the financial field. Indeed, the development of the Internet, and the increase of social networks users has led to the creation of a huge amount of data and information, which would be really useful to collect the different points of view of users. However, it is difficult

to automatically analyze data by directly reading a very large number of comments. Using transformer models make such a evaluation possible. A transformer network is an attention-based architecture for modeling sequential information : actually, transformers are processing sequential data, but they don't require processing in order (meaning that e.g. the beginning and end of a sentence can be processed in different orders). This makes it possible to parallelize the tasks and speed up the training duration. BERT is a ready-to-use pre-trained models, trained on huge datasets. By simply adding an additional output layer on the head of the transformers structure, these models can be used for specific tasks, e.g. for classification or question and answering tasks.

References

[1] Jenni L. Bettman, Stephen J. Sault, and Emma L. Schultz. "Fundamental and technical analysis: substitutes or complements?" In: *Accounting & Finance* 49.1 (2009), pp. 21–36. DOI: 10.1111/j.1467-629x.2008.00277.x.

[2] Antoine Bordes et al. "Translating Embeddings for Modeling Multi-Relational Data". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 2787–2795.

[3] Ines Chami et al. "Low-Dimensional Hyperbolic Knowledge Graph Embeddings". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 6901–6914. DOI: 10.18653/v1/2020.acl-main.617. URL: https://www.aclweb.org/anthology/2020.acl-main.617.

[4] Augustas Degutis and Lina Novickytė. "THE EFFICIENT MARKET HYPOTHESIS: A CRITICAL REVIEW OF LITERATURE AND METHODOLOGY". In: *Ekonomika* 93.2 (Jan. 2014), pp. 7–23. DOI: 10.15388/Ekon.2014.2.3549. URL: https://www.journals.vu.lt/ekonomika/article/view/3549.

[5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[6] Xiao Ding et al. "Deep Learning for Event-Driven Stock Prediction". In: *IJCAI*. 2015.

[7] Xiao Ding et al. "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1415–1425. DOI: 10.3115/v1/D14-1148. URL: https://www.aclweb.org/anthology/D14-1148.

[8] Xiao Ding et al. "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation". In: *EMNLP*. 2014.

[9] Eugene F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2 (May 1970), pp. 383–417. DOI: 10.2307/2325486.

[10] Katherine Beal Frazier, Robert W. Ingram, and B. Mack Tennyson. "A Methodology for the Analysis of Narrative Accounting Disclosures". In: *Journal of Accounting Research* 22.1 (1984), p. 318. DOI: 10.2307/2490713.

[11] Ziniu Hu et al. "Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 261–269. ISBN: 9781450355810. DOI: 10.1145/3159652.3159690. URL: https://doi.org/10.1145/3159652.3159690.

[12] Yankai Lin et al. "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *AAAI*. 2015.

[13] Tim Loughran and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of finance* 66.1 (2011), pp. 35–65.

[14] László Nemes and Attila Kiss. "Prediction of stock values changes using sentiment analysis of stock news headlines". In: *Journal of Information and Telecommunication* (2021), pp. 1–20.

[15] John R. Nofsinger. "Social Mood and Financial Economics". In: *Journal of Behavioral Finance* 6.3 (2005), pp. 144–160. DOI: 10.1207/s15427579jpfm0603_4. eprint: https://doi.org/10.1207/s15427579jpfm0603_4. URL: https://doi.org/10.1207/s15427579jpfm0603_4.

[16] Cheol-Ho Park and Scott H. Irwin. *The Profitability of Technical Analysis: A Review*. AgMAS Project Research Reports 37487. University of Illinois at Urbana-Champaign, Department of Agricultural and Consumer Economics, Oct. 2004. DOI: 10.22004/ag.econ.37487. URL: https://ideas.repec.org/p/ags/uiucrr/37487.html.

[17] Gupta S. Schneider M.J. "Forecasting sales of new and existing products using consumer reviews: A random projections approach." In: *International Journal of Forecasting* 32 (2016), pp. 243–256. DOI: 10.1016/j.ijforecast.2015.08.005.

[18] Sahar Sohangir, Nicholas Petty, and Dingding Wang. "Financial sentiment lexicon analysis". In: *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE. 2018, pp. 286–289.

[19] *Tracking Progress in Natural Language Processing*. http://nlpprogress.com/. Accessed: 2021-03-10.

[20] M. R. Vargas, B. S. L. P. de Lima, and A. G. Evsukoff. "Deep learning for stock market prediction from financial news articles". In: *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. 2017, pp. 60–65. DOI: 10.1109/CIVEMSA.2017.7995302.

[21] Frank Z. Xing, Erik Cambria, and Roy E. Welsch. "Natural language based financial forecasting: a survey". In: *Artificial Intelligence Review* 50.1 (2017), pp. 49–73. DOI: 10.1007/s10462-017-9588-9.

[22] Xi Zhang et al. “Improving stock market prediction via heterogeneous information fusion”. In: *Knowledge-Based Systems* 143 (2018), pp. 236–247. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2017.12.025. URL: https://www.sciencedirect.com/science/article/pii/S0950705117306032.