A two-step optimised BERT-based natural language processing algorithm

Rapheal Olaniyan, Daniel Stamate, and Ida Pu

Data Science & Soft Computing Lab, and
Computing Department, Goldsmiths, University of London

Abstract. Sentiment analysis involving the identification of sentiment polarities from textual data is a very popular area of research. Many research works that have explored and extracted sentiments from textual data such as financial news have been able to do so by employing BERT-based NLP algorithms in applications with high computational needs, and also by manually labelling sample data with help from financial experts. We propose an approach which makes possible the development of quality NLP models without the need for high computing power, or inputs from financial experts in labelling focused dataset for NLP model development. Our approach introduces a two-step optimised BERT-based NLP model for extracting sentiments from financial news. Our work shows that with little effort that involves manually labelling a small but relevant and focused sample data of financial news, one could achieve a high performing and accurate multi-class NLP model on financial news.

Keywords: Sentiment analysis · Financial news · NLP · Transfer learning · Classification · Two-step optimised BERT.

1 Introduction

The internet is full of online expressions which could be in the form of social blogs, financial news, or other kinds of textual expressions - thanks to the advancement in computer systems. With this advancement comes the ease of accessing, storing and processing large amounts of textual data. And now, sentiment analysis which helps to detect the element of feelings in textual data has become popular due to its vast applicability in areas such as artificial intelligence, stock market trading, politics, psychology, among others (Qie et al. [13], Bechara [2] and Hatfield [3]). For example, the polarity of sentiment extracted from textual data can be identified using sentiment analysis. This may be categorised as factual, positive reflecting a happy state of mind, negative referring to sad mood, or neutral. In addition, one may also use sentiment analysis to assess the degrees of polarised sentiments by scoring the different polarities of sentiments.

Sentiment analysis in the domain of finance, especially where the sentiments obtained are used to improve the predictive power of stock market predictive models, is of utmost importance. The predictive value of sentiments is highly time-sensitive with respect to first mover advantage in the face of market imperfection (Vayanos and Wang [11]). That is, one would expect the prices of the stock market to reflect all available information. As new information becomes available, players in the market adjust their positions and this new information becomes fully incorporated into the prices. There is a gap between these two points of information arrival and the time when prices reflect the new information. This short time window is termed as market imperfection (Vayanos and Wang [11]). This aspect supports the rationale behind the time sensitivity of the statistical significance of sentiment variables. Financial news could be the source of new information, expressed as sentiment, which has proved to be useful in enhancing stock market prediction with statistical and machine learning approaches (Smales [20], Shiller [25], Olaniyan et al. [26], and Marechal et al. [12]). Advanced NLP approaches are powerful tools that can be used to reliably and effectively extract sentiment polarity information from financial news, and we propose such a novel approach here based on adapting and extending the BERT algorithm [16].

Worryingly, it appears difficult - or so it seems - applying supervised NLP methods in this domain for these two obvious reasons: 1) Developing NLP classification models requires significant amount of effort to correctly label huge amount of training dataset to be used in the model training development. 2) The model to develop depends on domain-specific corpus for learning transfer as opposed to any general corpora which are not well-suited to supervised tasks.

As a result of these concerns NLP transfer learning methods have become a popular choice. They have been proven to be very promising and advanced the state of the art across natural language tasks and the foundation of these models, language model pre-training, is considered effective as the initial step required when developing natural language models (Dai and Le [4], Dolan and Brockett [28], Howard and Ruder [18], Baevski et al. [1]). The rationale for this choice of models is that they learn contextualized text representations by predicting words based on their contexts using very large corpora, and can be fine-tuned to adapt to downstream tasks (Peters et al. [21]). The challenge from the paucity of labelled data is avoided as the LM does not depend on it - rather it predicts words from contexts based on the semantic information it has learnt. And the fine-tuning of the NLP transfer learning methods on labelled data uses the semantic information learnt to predict labels. Here is where the problem lies: the fine-tuning of the NLP model on reliable labelled downstream tasks.

Manually labelling data for fine-tuning is difficult. First, it requires much time and effort. Second, the manually labelled data must be reliable and representative. Table 1 presents the experimental results on Financial PhraseBank [14] of 4845 financial news that were randomly selected from LexisNexis database and annotated by 16 financial experts. Interestingly, all the participants were able to agree on just 46% of the data's sentiment polarities. This clearly confirms the inherent challenge in manually labelling financial news data and, as a result, fine-tuning of models would suffer as a consequence.

Table 1. This table was taken from Araci [8]. Distribtution of sentiment labels and agreement levels in Financial PhraseBank

Agreement level	*Positive*	*Negative*	*Neutral*	*Count*
100	%25.2	%13.4	%61.4	2262
75%-99%	%26.6	%9.8	%63.6	1191
66%-74%	%36.7	%12.3	%50.9	765
50%-65%	%31.1	%14.4	%54.5	627
All	%28.1	%12.4	%59.4	4845

Considering these challenges our research focus is centred on developing a reliable NLP model for sentence polarity identification of financial news and we aim to achieve this with very minimal effort. In fact, our proposed two-step optimised BERT-based model overcomes these challenges. The main contributions of this work are therefore highlighted below:

1. We propose a two-step approach that includes 1) a primary model that relies on the labelled data from the experiment results on Financial PhraseBank and an optimised BERT-based NLP model called Roberta NLP, and 2) a secondary model that combines the experimental results and a small sample of financial news that have been manually labelled by us and validated with the primary model. This is to ensure that the secondary model has been fine-tuned with focused data.
2. We evaluate the primary model, compare with other related works in terms of their respective degrees of accuracy. The aim of this comparison is to see how the model fine-tuned with just the experimental results would perform on the financial news data related to the constituents of the S&P 500 index. The data is obtained from Intrinio platform [15].
3. We evaluate the results of the secondary model and compare with the primary model with the aim of assessing if the results obtained from both models are statistically different. Findings from the results would help us to understand the relevance of the secondary model especially when they are trained on focused data. In addition, we aim to assess the quality of our proposed two-step BERT-based model that does not rely on high computing machine and inputs from financial expert in manually labelling focused data for fine-tuning.

The rest of this research work is structured as follows: Section 2 presents the basis of the NLP model that would be used in this work. Section 3 provides information about the data sources, the various datasets, and the methodology applied. Results from the use of the primary model is presented in Section 4. Section 5 details the rationales behind the use of the secondary model and presents some empirical findings about why some models with very high level of accuracy may perform poorly on real life data. Finally, section 6 provides the conclusion to our work.

2 BERT NLP

We use the BERT(Bidirectional Encoder Representations from Transformers) model introduced by Devlin et al. [16]. As the name implies the centre-piece of the BERT model is transformer which was first published by Vaswani et al. [6]. Transformer is a great breakthrough in the world of language modelling.

The likes of convolutional neural networks (CNN) and Long-Short Term Memory (LSTM) are also good in language modelling but there are some constraints around them. One of these constraints is their poor performance when it comes to processing long sentences - the probability of learning the contextual relations between words when they are far away from each other diminishes linearly (Kalchbrenner et al. [23]) or exponentially (Gehring et al. [17]) depending on the language model used. Although some transduction models -models that convert input sequence of elements into another output sequence - have been able to overcome this challenge through the coupling of neural nets with attention learning mechanism that facilitates attention learning of specific words with the notion that these words could be embedded with contextual relevance (Bahdanau et al. [9], Kim et al. [29], Parikh et al. [5]). Another common problem with these transduction models is their inflexibility to parallel computation of tasks or their inefficient flexibility to it. This is where transformer plays the leading role whereby it does not depend on any coupling of neural nets with attention mechanism. It uses its inherent self-attention mechanism solely to draw the contextual relations between input and output and it also allows for efficient parallel computation of input and output (Devlin et al. [16]).

In the wake of the transformer many language model pre-trainings have sprung up and results from research works do support that these models are effective for enhancing NLP-related undertakings (Peters et al. [21], Howard and Ruder [18]). These models are applicable in a broad range of tasks such as named entity recognition (Li et al. [19]), sentiment analysis(Sun et al. [7]), text summarisation (Miller [10]), among others.

Most of the pre-training-based models are unidirectional - from the left to the right - in learning the general language representations. Devlin et al. [16] state that such architectural constraint limits the choice of architectures in the first place and are sub-optimal for sentence-level tasks. In view of this BERT is proposed for fine-tuning because of its uniquely bidirectional approach for general language representations.

The optimised version of the pre-trained BERT model will be used in this work as originally presented by Liu et al. [30] for developing the primary model. This model will be the basis upon which the secondary model is developed.

3 Methodology and Data

In the process of conducting sentiment analysis on financial news we source for financial news data from Intrinio platform [15]. The data collected covers the period September 2012 and July 2019 and we have 1.05 million records -

our interest is to extract multiclass sentiment polarities from this data. The financial news data collected are related only to the constituents of the S &P 500 index. Extra care is required in ensuring that false positive and negatives are minimised. In doing this we propose a two-step optimised BERT-based NLP model where the first step is to produce the primary model that explores both the labelled data from the experimental results on Financial PhraseBank and the optimised BERT-based NLP model. The level of accuracy of the model would be examined to see if it qualifies enough to become our primary model. More specifically, we employ Roberta NLP model which is considered the optimised version of the BERT model. As mentioned earlier, two steps would be followed in the sentiment analysis process. First, we train the optimised BERT-based NLP model with the experimental results on the financial Phrasebank dataset which is the same dataset used in [24] and [8]. The dataset consists of 4,845 financial news that were randomly selected from the LexisNexis database. In the process of manually labelling the financial news data 16 financial professionals were asked to participate. 47% (2263 of the 4845) of the financial news had 100% agreements from all the participants. This implies that some sentences were assigned different labels by different participants. Clearly, this is a confirmation that manual labelling is complex and challenging to correctly assign true labels due to varyingly and subjectively contextual perspectives. It is also laborious to manually label high volume of sentences for developing training models.

In view of these issues we resolve to using only the financial news with 100% agreement level from all the participants totalling 2263 sentences. The summary of the selected sample data is presented in table 2.

Table 2. Experimental results on the Financial PhraseBank dataset containing the 2263 financial news with 100% agreement level labelled by the 16 survey participants with financial background.

Value	*Polarity*	Count
0	Neutral	1390
1	Positive	570
-1	Negative	303
Grand total		**2263**

The primary model is a key and integral part of the secondary model. Attention would therefore be paid to the degree of accuracy of the model with the assumption that a high degree would constitute its acceptance as a basis for developing the secondary model. Recall that 16 professionals were involved in the experimental labelled results on the Financial PhraseBank. One of the possible reasons for involving many financial professionals was to ensure that the experimental results produced were reliable and of high quality from the labelling task.

Understandably, going through at least the same level of effort of involving many financial experts in the manual labelling is resource-consuming and time-taking. As a result, we are proposing a two-step approach that we consider to be effective both in labelling and in model training. It is worth mentioning that the manual labelling of sentences itself is challenging not just in regard to the volume of sentences to label but also in correctly labelling sentences because sometimes there seems to be a very thin line among the classes e.g. positive and neutral for example.

Below are some examples:

1. InvestorPlace Stock Market News Stock Advice amp Trading Tips Apple NASDAQ AAPL will be reporting its third quarter earnings on July. Apple stock has performed well since the start of June posting a gain since June but on July all bets are off.
2. Why Apple Stock May Be a Case of Near-Term Pain, Long-Term Gain
3. UPDATE -Ireland invests disputed Apple taxes in low-risk bonds
4. American CEO reiterates confidence in Max return by mid-August despite unclear timetable from Boeing, FAA
5. UPDATE -Apple explores moving -% of production capacity from China - Nikkei

Manually labelling over 1 million financial news would be very laborious and we would expect a lot of disagreements in labelling some news among us: hence, the need to manually label a small sample and use the trained model to validate the results of our labelled news. Where we have disagreements in the results between our manual labelling and the trained model, we review carefully in order to identify the true labels. We are more interested in the false positives and negatives from the model's results so that we could review the sentences that we consider to be wrongly labelled and add the reviewed sentences to the training data and finally obtain the secondary model. Eventually with help from the trained primary model, we have 2,000 labelled news - that have been randomly selected from [15] - to be added to the original training data. Our secondary model is therefore trained by combining the 2000 labelled news that have been reviewed and the experimental results on the financial Phrasebank. The trained secondary model is then applied to over 1 million financial news data in order to obtain sentence polarities which could be positive, neutral or negative.

4 Primary NLP model

Before the BERT NLP model could be used it has to go through two key steps. First, the BERT would have to be pre-trained like every other language model so that they could learn the contextual relations between words. Pre-training a model on a very large corpus is a very resource-consuming effort especially with the amount of time and computing architecture capacity required. For example, most of the BERT pre-training exercises were conducted on the Google cloud([16], [31]) and Amazon cloud([8]) translating to the high dependence of the pre-training stage on high computing machines. Devlin et al. [16] pre-trained the model using the corpus that contained the combination of the BooksCorpus (800M words) (Zhu et al. [31]) and English Wikipedia (2,500M words).

The second stage would require that the pre-trained model goes through supervised learning where the training dataset contains texts and their respective labels e.g. "US stock market is bullish" is the text and the label is "positive". The results predicted using the trained BERT-based models have been promising and this accounts for its popularity.

Araci [8] examined if by both pre-training(unsupervised learning) and training (supervised learning) the BERT on downstream tasks could improve further the BERT model. In the process, the author pre-trained the BERT model on the financial news data obtained from Reuters at first, and then trained the model using the experimental results on the financial Phrasebank dataset which was the same data used in [24]. Findings showed that such process could improve the model performance by 15% in accuracy.

Liu et al. [30] revisited the work done by [16] and concluded that the pre-trained BERT was not at its optimal level. They pre-trained the model all over and finally obtained the optimised BERT model. In view of this development we would be using the optimised BERT model to compare the results with the Araci [8]'s. Findings from this work would help us to answer the following questions:

1. Is pre-training the BERT model with targeted downstream task necessary as opposed to the huge corpus of BooksCorpus (800M words) and English Wikipedia used to pre-train the BERT?
2. Is the downstream data used by [24] and [8] for the supervised model training and evaluation representative enough of the financial news?

Table 3. Most of the information in this table was taken from Araci [8] with reference to Malo et al. [24], Krishnamoorthy [27] and Maia et al.[22] regarding the results using LPS, HSC and FinSSLX respectively. The last row is added to the table and it represents the results obtained from the optimised BERT-based model - this report was based on a 5-fold validation results.

Data with 100 % agreement			
Model	Loss	Accuracy	F1 Score
LSTM	0.57	0.81	0.74
LSTM with ELMo	0.50	0.84	0.77
ULMFit	0.20	0.93	0.91
LPS	-	0.79	0.80
HSC	-	0.83	0.86
FinSSLX	-	0.91	0.88
FinBERT	0.13	0.97	0.95
ROBERTA	**0.12**	**0.97**	**0.97**

As shown in table 3 the optimised BERT (ROBERTA) model appears to have achieved the same level of accuracy as the FinBERT model. The optimised BERT is only trained with the downstream tasks and the results show that it performs as highly accurate as the FinBERT's which was proposed by [8]. In view of this it would be hard to conclude that pre-training the BERT model with downstream tasks would improve the accuracy level of the BERT as [8] would have claimed.

The optimised BERT model performs well with a very high level of accuracy when trained on the Financial PhraseBank dataset. Would this trained model perform well on real life data? To answer the question, we apply our trained model on a new financial news data in order to evaluate its reliability when applied to real life situation. This task is addressed in the next section.

5 Secondary NLP model

We use the experimental results on the Financial PhraseBank data to train the optimised BERT model and this achieves a high accuracy level of 97% as presented in table 3. In order to assess how representative the training data is, we evaluate the predicted results obtained from the trained model (primary model) on a new set of financial news obtained from [15]. If the results obtained show at least 90% level of accuracy on the new data, we would conclude that the training data is highly representative of the financial news and that the model seems reliable without the need for the proposed secondary model.

In the process we start by manually labelling 3000 financial news randomly sourced from [15]. We understand the concern that we might not be 100% accurate in the manual labelling; hence, the need to rely on the primary model for the validation of the manual labels and the review and correction of the labels that appear as false positives and false negatives.

Table 4. This report represents the out-of-sample confusion matrix between the results from the validated manual labelling and the primary model. The results show 82% level of accuracy.

Class	Precision	Recall	F1 Score	Support
-1	0.93	0.93	0.93	42
0	0.86	0.59	0.70	63
1	0.77	0.93	0.84	91

The results from the model is presented in table 4. The results show that when the trained primary model is applied to predict the sentiments from a different sample data the level of accuracy drops significantly to 82%. On this ground we conclude that the initial sample data - the experimental results on the Financial PhraseBank dataset - is short of being considered as a representative of the financial news in general. Considering this we develop the secondary model which is the second step of our prosed two-step optimised BERT-based NLP model. The secondary model clearly shows improvement judging by the accuracy level. The test is applied on a series of sample datasets and they all show promising results. With series of tests we are able to achieve accuracy level that is at least 97%. One of the results is presented in table 5.

Table 5. This report represents the out-of-sample confusion matrix between the results from the manual labelling and the Secondary model. The results show 99% level of accuracy.

Class	Precision	Recall	F1 Score	Support
-1	0.98	1.0	0.99	42
0	1.0	1.0	1.0	63
1	1.0	0.99	0.99	91

6 Discussion and conclusion

The use of BERT models in NLP modelling has gained huge popularity and we have been able to demonstrate the promising results from its applicability. Many works have been developed using the BERT family. Some have attempted to improve on the BERT-related work by proposing that pre-training the BERT models on downstream tasks would provide improved results from such an endeavour (Araci [8]). This suggested approach is clearly laboriously resource-consuming in that one would have to source for pre-training downstream tasks and that inputs from experts would be required in manually labelling sample data for model fine-tuning.

We acknowledge how arduous manually labelling a high volume of financial news data for training NLP models could be and the need to involve financial professional experts in the labelling effort. In view of these challenges we have proposed a two-step optimised BERT-based approach which has the tendency of minimising the impacts of not including these two requirements. That is, with our proposed approach we would need to manually label just a small sample of focused financial news data. And noting that our manually labelled training data might contain some false labels our approach has therefore taken into account such occurrence by searching for false labels (otherwise known as mismatches between the results from the primary model and our manual labels) and correcting before they are fed into the secondary model which is the second step of our proposed model. This is done by establishing a primary model developed by first training the optimised BERT model using the experimental results from the Financial PhraseBank news data. The trained primary model is then applied to the small sample of 3000 manually labelled news for validation. In doing so false matches between the manual labelling and the results from the primary model are identified and reviewed. The manually labelled data - reviewed and true matches - are then added to the initial training data resulting in a combined training data which would then be used to re-train an optimised BERT model. This becomes the trained secondary model. The rationale for these two steps is that the trained primary model on its own is not sufficient due to the misrepresentation of the initial training data as shown by its low accuracy level of 82%. But with the secondary model that has been developed with more focused data and little manual inputs we are able to achieve a high accuracy of over 97%. This is achieved with a minimal and manageable efforts.

References

1. A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli: *Cloze-driven pretraining of self-attention networks*,arXiv preprint arXiv:1903.07785, 2019.
2. A. Bechara: *The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage*, Brain and Cognition,Development of Orbitofrontal Function, 55(1):30 – 40, 2004.
3. A. Hatfield, J. Cacioppo, and R.L. Rapson: *Emotional contagion*, Cambridge University Press, 1994.
4. A.M. Dai and Q.V. Le: *Semi-supervised sequence learning*, In Advances in neural information processing systems, pages 3079–3087, 2015.
5. A. Parikh, O. Täckström, D. Das, and J. Uszkoreit: *A decomposable attention model*, In Empirical Methods in Natural Language Processing, 2016.
6. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin: *Attention is all you need*, 2017, arXiv:1706.03762
7. C. Sun, L. Huang and X. Qiu: *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence*, arXiv:1903.09588, 2019.
8. D. Araci: *FinBERT: Financial sentiment analysis with pre-trained language models*, arXiv:1908.10063, 2019.
9. D. Bahdanau, K. Cho, and Y. Bengio: *Neural machine translation by jointly learning to align and translate*, CoRR, abs/1409.0473, 2014.
10. D. Miller: *Leveraging BERT for extractive text summarization on lectures*, arXiv:1906.04165, 2019.
11. D. Vayanos and J. Wang: *Market Liquidity – Theory and Empirical Evidence*,Handbook of the Economics of Finance Volume 2, Part B, Pages 1289–1361, 2013.
12. F. Marechal, D. Stamate, R. Olaniyan and J. Marek: *On XLE Index Constituents' Social Media Based Sentiment Informing the Index Trend and Volatility Prediction*, Proceedings of the 10th Intl. Conference on Computatational Collective Intelligence (ICCCI), Springer, LNCS, 2018.
13. G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu and C. Chen: *DASA:Dissatisfaction-oriented Advertising based on Sentiment Analysis*, Expert Systems with Applications,37(9):6182–6191, 2010.
14. https://www.researchgate.net/publication/251231364_FinancialPhraseBank–v10
15. https://intrinio.com/
16. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018, arXiv preprint arXiv:1810.04805.

17. J. Gehring, M. Auli, D. Grangier, D. Yarats and Y.N. Dauphin: *Convolutional sequence to sequence learning*, arXiv preprint arXiv:1705.03122v2, 2017.
18. J. Howard and S. Ruder: *Universal Language Model Fine-tuning for Text Classification*,arXiv:1801.06146 http://arxiv.org/abs/1801.06146,2018.
19. J. Li, A. Sun, J. Han, C. Li: *A Survey on Deep Learning for Named Entity Recognition*,arXiv:1812.09449, 2018.
20. L. A. Smales: *Non-scheduled news arrival and high frequency stock market dynamics: Evidence from the Australian Securities Exchange*, Research in International Business and Finance. vol. 32, pp. 122–138, 2014.
21. M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer: *Deep contextualized word representations*, 2018 from https://doi.org/10.18653/v1/N18–1202 arXiv:1802.05365.
22. M. Maia, S. Handschuh, A. Freitas, B. Davis, R. Mcdermott, M. Zarrouk, A. Balahur and R. Mc-Dermott: *Companion of the the web conference 2018 on the web conference 2018*, WWW 2018, Lyon , France, 23–27, 2018.
23. N. Kalchbrenner, L. Espeholt, K. Simonyan, A.V.D. Oord, A. Graves and K. Kavukcuoglu: *Neural machine translation in linear time*, arXiv preprint arXiv:1610.10099v2, 2017.
24. P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala: *Good debt or bad debt: Detecting semantic orientations in economic texts*,Journal of the Association for Information Science and Technology, 65(4):782–796, 2014.
25. R.J. Shiller: *Irrational Exuberance* Princeton: Princeton University press, 2000.
26. R. Olaniyan, D. Stamate, D. Logofatu and L. Ouarbya: *Sentiment and Stock Market Volatility Predictive Modelling - a Hybrid Approach*, Proceedings of the 2nd IEEE/ACM International Conference on Data Science and Advanced Analytics, 2015.
27. S. Krishnamoorthy: *Sentiment analysis of financial news articles using performance indicators*, Knowledge and Information Systems, 56(2):373–394, 2018.
28. W.B. Dolan and C. Brockett: *Automatically constructing a corpus of sentential paraphrases*,In Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005.
29. Y. Kim, C. Denton, L. Hoang, and A.M. Rush: *Structured attention networks*, In International Conference on Learning Representations, 2017.
30. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov: *RoBERTa: a robustly optimized BERT pretraining approach*, arXiv:1907.11692, 2019.
31. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler: *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.