MULTI-TASK LEARNING ON FinBERT FOR FINANCIAL FINE-GRAINED SENTIMENT ANALYSIS

Christelle George
11288106
Christelle.George@hec.ca
Dpt. of Decision Science

David Lemieux
11118064
David.3.Lemieux@hec.ca
Dpt. of Decision Science

Marc-Antoine Provost
11205859
marc-antoine.provost@hec.ca
Dpt. of Decision Science

1 Introduction

Financial markets are looked at with awe and confusion by market participants, investors, data scientists and anybody that tries to forecast prices or extract valuable information in order to invest in such markets. The constant flow of information, let it be from news or analyst reports, have a direct impact on investors' positions and market prices. However, vast amount of new financial information is created every day and manually deriving actionable insights from it is impossible for any single entity. Thus, automatic sentiment analysis of texts becomes critical for firms and investors as it provides faster access to salient information and can give actors an edge. In fact, automatically inferring the polarity of a financial news or headline saves time and provide valuable insights. Even though massive amount of new financial data is created everyday, there is a serious lack of labeled financial data. Moreover, it is largely domain specific as opposed to traditional news data and require heavy domain knowledge to decipher. Hence, traditional deep learning models such as Recurrent Neural Networks as well as state-of-the-art data-hungry text classification models are quite ineffective in regards to financial sentiment analysis. As such, the principal research interest for this thesis is polarity analysis. More specifically, we investigate two areas of research related to Natural Language Understanding in the field of finance. We heavily base our work upon previous research, namely the FinBERT paper by [Ara19] and described in section 2.

We first hypothesize that a MultiTask learning approach offers better performance than a single-task approach on financial sentiment analysis as proposed by the FinBERT authors. Furthermore, we hypothesize that the unfreezing strategy of the pre-trained layers as stated in the FinBERT paper should be part of the hyper-parameter tuning step since it is expected to increase the model's performance. Thus, the goal of this thesis is to test the hypothesized advantages of exploring different unfreezing strategies as well as using fine-tuned pre-trained language models for financial domain in a Multi-Task scenario.

1.1 Contribution

The highlights of important research in the field, the analysis of previously mentioned hypothesis, the methodology and structure used for the construction and tuning of the models allow us to target three main contributions to the research community. First, the use of a Multi-Task Learning approach with the FinBERT model was unheard of. Secondly, the analysis of the unfreezing strategy is also not common in the literature and was never done with the FinBERT model. We lastly contribute on a more technical and educational basis by going into detail in our hyper-parameter tuning done using the Weight & Biases platform. This is specifically interesting since few research papers actually go in depth in this critical step of building efficient deep learning models.

1.2 Structure

The rest of the thesis is organised as follows: First, relevant literature in sentiment analysis and pre-trained language models are discussed in section 2. Then, a detailed explanation of the methodology employed is provided in section 3. This is followed by a presentation on experimental results on the

financial sentiment dataset 4. Finally, we conclude by providing further research ideas in order to improve upon our work.

We invite the interested reader to access the code, data and hyper-parameter tuning platform at `https://github.com/dlemieux89/Deep_Learning2` and `https://wandb.ai/quantolio_dle/DL2_FinBERT_1task?workspace=user-quantolio_dle`.

2 Literature Review

Modern sentiment analysis is an interesting topic that spurred with the arrival of internet reviews in the mid 2000's [VGK16]. Since then, the use of sentiment analysis has reached numerous domains, such as media and finance. Financial sentiment analysis differs from traditional sentiment analysis in domain and purpose. In fact, where general sentiment analysis has been about people's opinion polarity in regards to a product or service, financial sentiment analysis is about predicting the polarity of the financial markets given textual data. One of the first work tackling this problem using deep learning is from [KF17], where the authors used a Long-Short-Term-Memory (LSTM) network with the idea of transfer learning. LSTM networks are a variant of Recurrent Neural Networks (RNN) that excel at processing sequential data. The introduction of forget gates in LSTM networks alleviates the exploding and vanishing gradient problem that hindered learning long-term dependencies with standard RNNs. However, due to the lack of large labeled financial datasets, such models are often not used to their full capacity. As such, researchers have found that training a model on a very large corpora and then fine-tuning it on a downstream task can be a very efficient strategy. In fact, this greatly helps with the scarcity of labeled data since language models' task is to predict the next word. This concept has been around for quite a few years in Natural Language Processing (NLP), but was largely publicized with the Embeddings from Language Models (ELMo) [Pet+18] paper. The authors showed that pre-training a deep bi-directional language model a on large text corpus and using the hidden states to learn contextualized words representations significantly improved performance on most NLP tasks. In fact, initializing embeddings for downstream tasks with the deep contextualized word representations learned from ELMo was shown to perform better than static word embeddings methods such as word2vec [Mik+13] or GloVe [PSM14].

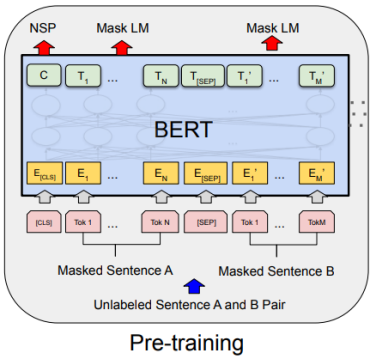

Figure 1: BERT Pre-training

The pre-training approach was taken further with the introduction of Bidirectional Encoder Representations from Transformers (BERT) [Dev+18]. Its arrival announced a new era for NLP by breaking several records and by making its code publicly available. The innovations brought by BERT come from its definition of the language modeling task and its training on unprecedentedly large corpus. The approach consists of first pre-training the model using a combination of Masked Language Modelling (MLM) and a Next Sentence Prediction (NSP) task (refer to figure 1). Following the pre-training phase, the model can be fine-tuned on the task of interest.

The authors of FinBERT based their work BERT to tackle financial sentiment analysis [Ara19]. Indeed, they conducted further pre-training of BERT on financial corpus and explored some training strategies to prevent catastrophic forgetting, such as gradual unfreezing. Rather than fine-tuning all layers at once, gradual unfreezing proposes to gradually unfreeze the model starting from the last layer as it contains the least general knowledge [HR18]. With these design choices, FinBERT was able to achieve state-of-the-art performance on financial sentiment analysis.

Another approach that has gained popularity in the NLP community is multi-task learning. It is a field of machine learning where multiple tasks are learned in parallel while using shared representations [PCL20a]. The intuition behind this method is that if the tasks are correlated, the learner can jointly learn a model for them while taking into consideration the shared information, which is expected to improve its generalization ability. As people express their opinion on various subjects with different styles, a multi-task approach can be relevant. Specifically in the financial domain, the different settings of classification like binary and ternary are correlated since their difference lies in the sentiment granularity of the classes which increases while moving from binary to fine-

grained problem. In regards to sentiment analysis, Balikas and Al. [BMA17] explored this method for fine-grained Twitter sentiment analysis, where the tasks were ternary and fine-grained sentiment classification. They showed that by jointly learning the tasks with a multi-task learning model such as a RNN, one can greatly improve the performance of the second task. Moreover, building upon multi-task research, an empirical study of multi-task learning with BERT was conducted on multiple biomedical and clinical NLP tasks [PCL20b]. The authors demonstrated that a multi-task learning approach to obtain a single model achieved state-of-the-art performance over a single-task model on all tested tasks. This work validates the soundness of our hypotheses and sheds light on important algorithmic design decisions.

3 Methodology

In this section we highlight the methodology used to investigate the two hypotheses postulated previously. The following subsection discusses the data (3.1), the modeling (3.2), the evaluation metrics (3.4) and finally the hyper-parameter tuning (3.5).

3.1 Data

The data used is called "Financial PhraseBank". It was created by [Mal+14]. The dataset has 4845 english sentences selected at random from financial news on LexisNexis database. The 3 labels (positive, neutral, negative) were carefully selected by 16 experts with a background in finance. It is important to mention that the given labels do not correspond to the feelings in a traditional sentiment analysis, but rather to the effect the information can have on the company's stock price. In other words, the negative sentiment is associated to a financial loss; the positive sentiment to a gain. It is important as well to keep in mind that the labels are subjective, and might include a bias in our analysis. Also, the dataset includes information on the level of agreements on sentences by the expert. Finally, the dataset has a class imbalance : almost 60% of all sentences are neutral. It is possible to see some examples for each class in the list below.

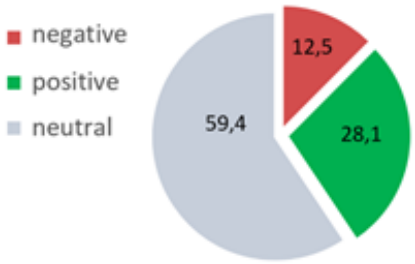

Figure 2: Data Dispersion for the 3 classes

- Positive : Net sales increased by 5.2% to EUR 205.5 mn, and operating profit by 34.9% to EUR 23.5 mn
- Neutral : Deliveries are to start later in 2010, and the volume will increase in the years 2011-2012
- Negative: Compared with the FTSE 100 index, this was a relative price change of -0.4% Deliveries

3.2 Models

To investigate and answer the first previously exposed hypothesis we propose the construction of two benchmark models that will be compared with the Multi-Task FinBERT model. The first benchmark model is the construction of an RNN variant, a Gated Recurrent Unit (GRU) [Chu+14] model. It is a variant of LSTM Networks, where the main difference is the use of LSTM as hidden units, as they proved their capability for explicit memory deletes and updates. Such model was implemented in Pytorch and pre-trained GloVe embeddings [PSM14] from Wikipedia and news articles were used with a fixed dimension of 300. As for other hyper-parameters, 128 hidden neurons and 1 hidden layer were used and trained for 25 epochs.

The second and presumably more competitive benchmark used was the original FinBERT model as per the structure directly found on HugginFace (`https://huggingface.co/`). In this structure, the level of Hyper-Parameter Tuning (thereafter HPT) is somewhat restricted to the variation of the Learning Rate (LR) and other parameters such as the Unfreezing strategy is left constant with a custom descending pattern where the layers are unfreezed in the order 10,8,6,4,2,0 for each subsequent epoch for each of the FinBERT models.

This original FinBERT model is not optimized to have the best performance but rather constructed to give an idea of the performance of this category of model on the dataset. While it is true that we

could have selected the model to be fully frozen, this was not chosen since we wanted to allow more flexibility and generability for the model.

After the 2 benchmark models were created and lightly tuned the Multi-Task model was created and compared. Because this paper is at the intersection of a Scientific article and an Academic Report, we will go into further details than necessary in the former category of publication. The goal being to demonstrate what was done in details, in order to allow the reader to understand the subtleties applied in the current work.

Finally, to evaluate the second hypothesis, the best learning rate from the previous best original FinBERT model will be used. Here, we propose to analyse multiple Unfreezing strategies and compare them altogether. The assumption here conveys that a more adapted structure can provide extra performance compared with no freezing strategies at all. To test the different freezing strategies we propose to look at the following:

- No freezing
- All freezing
- Ascending freezing
- Descending freezing
- Random layer freezing (3)

When researchers use pre-trained models, it inherently comes with some weighed parameters embedded in the trained model. When we train a pre-trained model without freezing the first couples of layers it tries to change the weights to adapt them to the newly seen piece of information. However, the pre-train model such as BERT, RoBerta and FinBERT are trained on millions of lines of words and the embeddings that the model starts-with are robust. Because the model already starts with a strong structure of learned weights the way we allow the model to modify those weights is crucial.

3.2.1 Structure and Setup of the Multi-Task Model

This subsection specifically covers the structure and setup used for the Multi-Task Model discussed in previous section (3.2). The first step to the creation of the Multi-Task Model was modifying the original FinBERT model. From the article [Ara19] it is possible to see that the original model includes 12 encoding layers and 1 classification layer. In order to adapt the model to our current need the Sequence Module of layers were modified using the Pytorch [Pas+19] interface to remove the last classification and last pooling layer. The original classification problem was a 3 way classification and since our hypothesis includes a Multi-Task Fine-Grained approach the task 1 is only a two step class classifier. Below is a flow chart presenting the structure of the Multi-Task Model.

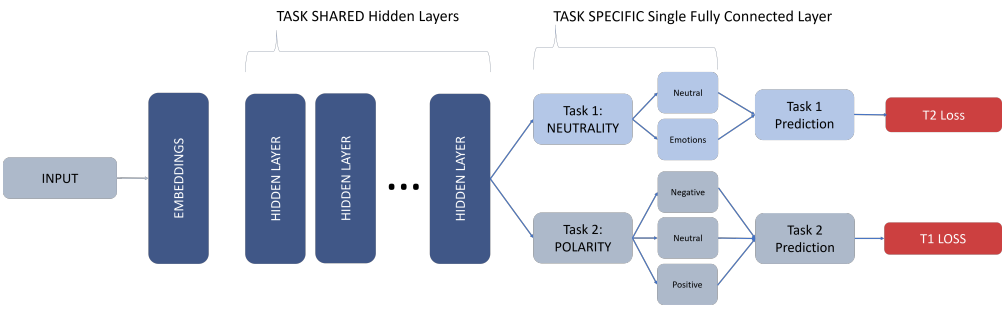

Figure 3: Structure for the Multi-Task model

It is possible to see, from the previous graph, the multiple heads stemming from the Task-Shared Hidden Layers that create 2 different Task-Specific fully connected layers. The first fully connected layer is specific for task one with a specification of 2 different outputs (Neutral or Emotion) and the second task that outputs a 3 way classification.

3.3 Training and Train-Test-Split

With the option stratify to correctly ensure a realistic split of the data, we choose to train-valid-test split our data with a 80%, 10% and 10% proportion. For the training process it is important to understand that the Multi-Task Learning Structure that we have built is not a Hierarchical or a

Parallel Learning Structure. In the present case the Multi-Task Head outputs the prediction of the task 1 (Emotion or Neutral) or Task 2 (Positive, Negative, Neutral) as per the flow chart in the previous section 3.2.1. The particular training process used can be described as follows.

For every batch in the training set, the task to train is either set to task 1 or to task 2. The selection is made using a uniform random number with a threshold arbitrarily set at 0.50. The goal is to allow the model to train on both task assuming that the training on task 1 and task 2 will benefit each other. Such method is also used in the work from [PCL20b]

In the current experiment, task 1 and task 2 are trained on the same dataset with different possible set of labels but other unrelated dataset could have be used. As stated previously, this type of dataset is rare and multi-dataset Multi-Task Learning was out of question for the current case.

3.4 Evaluation Metrics

For both hypothesis, we evaluated all models with two common metrics : Accuracy and F1 score. The accuracy gives an indication of how well the classification operated. However, it might not reflect the whole information, specially in the context of imbalanced classes. In that case, we also used the F1 score, which represents an weighted average of the precision and recall. This could shed to light a specific class being incorrectly classified. In this study, we use both the macro average F1 score, as well as the F1 score specific to each class (positive, negative, neutral).

3.5 Hyper-Parameter Tuning

This section not only set the environment for the analysis of hypothesis 2 on the unfreezing strategy, it wishes to highlight the methodology used for the global Hyper-Parameter Tuning experience. While Hyper Parameter tuning is rarely an explicit center piece in academic papers we will do a small caveat to discuss this briefly and introduce an interesting platform (`https://wandb.ai/`) The goal of this caveat is also to allow interested party to learn from this section as this paper is also created on the context of a academic course that is also to be presented at a "course level conference"

4 Results

4.1 Experiments

Studying the effectiveness of our Multi-Task Learning FinBERT using the Financial Phrase Bank dataset, we highlight the following results, confirming our initial hypotheses :

1- The Multi-Task Learning approach on FinBERT outperforms all 3 models/methods with both performance metrics tested.

2- The unfreezing strategy, being part of the hyper-parameter tuning step allows us to reach a better performance.

The performance of all four models is presented in Table 4 and 5 for hypothesis 1 and in table 6 for hypothesis 2 .

The benchmark GRU performs poorly, as the financial vocabulary is not captured by traditional models. In other words, the RNN models associates the positive sentiment to a happy thought, and the negative sentiment to an angry or sad sentiment. As we mentioned previously, the financial standpoint treats the sentiment with a gain or a profit connotation, which the RNN fails to capture. The F1 score associated to the negative and positive label is mediocre, with respectively 1,2% and 3,4%. The F1 score of 45,3% associated to the neutral label is not an indicator of a better performance, since the majority label is the neutral one: the model predicts almost all observations to the neutral class. Finally, the accuracy of 62,3% is clearly misleading as it reflects mostly the performance of the neutral class which represents around 60% of the data.

On the other hand, the hyper-parameter tuning on the FinBERT has showed its efficiency : it had increased by 1,92% the accuracy and by 2,75% the F1 score macro average, compared to the traditional FinBERT model. This highlight the importance of choosing optimal parameters, such as the optimizer AdamW, a learning rate of 0,0003 and a batch size of 64.

Last but not least, the Multi-Task learning approach used with the FinBERT achieved the best performance amongst all models, according to all metrics. By dropping the output layer of the FinBERT, using a multi head and training the model with the two tasks (neutrality and polarity) improved its

performance. This joint training has improved the F1 score of every label, and the best improvement goes to the minority class : the negative label, representing roughly 12% of the data, reached a 83,1% F1 score. Overall, the Multi-Task FinBERT scored a 89% accuracy, which represent a 43% improvement compared to the benchmark.

Model	Accuracy	F1 score		
		Negative	Neutral	Positive
GRU - Benchmark	62,30	1,20	45,30	3,40
FinBERT - Paper	86	84 (Global)		
FinBERT - Single Task	87,92	80,89	88,62	90,74
FinBERT - Multi Task	88,96	83,07	89,43	91,42

Figure 4: Performance for Multi-Task models

In this section, we briefly compare the single task versus the Multi-Task FinBERT. The confusion matrix illustrates the normalized performance of the Multi-Task FinBERT, whereas the annotations denote the differential between both FinBERT models. Both true negative and true positive labels have a positive differential since they represent the minority classes that were best predicted with the Multi-Task. The true neutral class has a negative differential of value 1, which is acceptable since it got less errors in terms of false neutral : its F1 score increased to 89,43%.

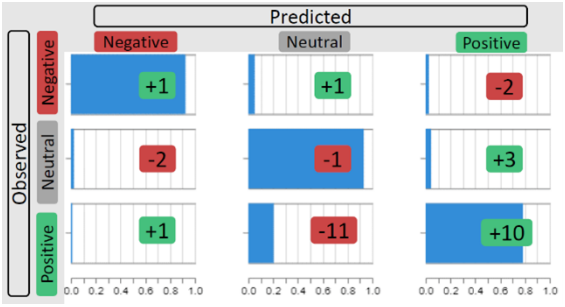

Figure 5: Confusion Matrix for the Multi-Task Model

The freezing approach revolved around 7 strategies. The strategy that reached the best performance is the one where all the layers were frozen, with a 89,17% accuracy and a 88,02% F1 score macro average. This comes with no surprise since the FinBERT model was pre-trained on the Financial PhraseBank dataset used: : the weights of each layer are already optimal. The descending freezing and all 3 random layer freezing strategies obtained a good performance that almost reached the optimal freeze-all strategy. Finally, the ascending freezing and no freezing strategy both got terrible results : a 59% accuracy with a 25% F1 score macro average. This is what is commonly known as catastrophic forgetting : unfreezing the first layers of the word embeddings leads to a change in their weights. The model no longer recognizes the vocabulary it was trained on and loses its carefully trained representation. As a result, it is no longer able to correctly predict the labels.

Strategy	Accuracy	F1 Score
Freeze all	89,17	88,02
Descending freezing	87,60	86,88
Random layer freezing 1	88,33	85,21
Random layer freezing 2	86,46	87,11
Random layer freezing 3	87,29	85,89
No freezing	59,58	24,88
Ascending freezing	59,58	24,88

Figure 6: Performance for the Defreezing strategy models

5 Conclusion

In this academic paper we investigated 2 hypotheses. More specifically we investigated if the FinBERT model with Multi-Task Learning could create extra performance compared to the original FinBERT. Using the same FinancialPhrase Bank dataset we have shown that keeping everything else constant the Multi-Task Learning model was able to have a better performance than the initial model. The literature on the Multi-Task approach would suggest the same conclusion but never before, to our current knowledge, this was done.

The second hypothesis we investigated is oriented with the technical aspect of unfreezing the layer of a pre-trained model. In most scientific paper the unfreezing aspect of the fine-tuning of a model is usually left for the authors or only lightly discussed. Some paper discussed the use of multiple unfreezing approach, it still remains a more obscure topic. In this paper we advance that the selection of the unfreezing strategy should be an important part of the learning in the field of deep learning just as much as the concept of learning rate. In this paper we have shown that the tuning of the unfreezing strategy has an impact on the performance. It shows that the absence of freezing is as detrimental as the wrong (ascending) defreezing strategy.

Finally, we briefly discussed hyper-parameter tuning and the platform Weights & Biases where HPT could be done with ease. We believe that the Hyper Parameter search is too often silenced in important academic paper and some results seem to be coming from arcane places. This little caveat for HPT in this paper is our advocacy for the transparency in the field of Model Tuning in Deep Learning.

Even though the work accomplished allows to innovate in the field of NLP, finance, and Deep Learning, it is only a small step in the right direction. To further enhance the work done in this paper it would be possible to improve in multiple subsections. In the data section it could be interesting to try some form of data augmentation. It is a well known fact that labeled data is a rare resource in finance. Data Augmentation could be a solution to the current problem and enhance the current performance. Also in this subsection, it would be possible to use other features such as the count of certain terms, the presence of emoticons (smiley faces) and number of capital letters to predict emotions. On the structural aspect of the current work, it would also be interesting to investigate Simultaneous Learning and Hierarchical Learning instead of Alternative Multi-Task Learning.

References

[Mik+13] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv e-prints*, arXiv:1301.3781 (Jan. 2013), arXiv:1301.3781. arXiv: 1301.3781 [cs.CL].

[Chu+14] Junyoung Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv e-prints*, arXiv:1412.3555 (Dec. 2014), arXiv:1412.3555. arXiv: 1412.3555 [cs.NE].

[Mal+14] Pekka Malo et al. "Good debt or bad debt: Detecting semantic orientations in economic texts". In: *Journal of the Association for Information Science and Technology* 65.4 (2014), pp. 782–796.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[VGK16] Mika Viking Mäntylä, Daniel Graziotin, and Miikka Kuutila. "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers". In: *arXiv e-prints*, arXiv:1612.01556 (Dec. 2016), arXiv:1612.01556. arXiv: 1612.01556 [cs.CL].

[BMA17] Georgios Balikas, Simon Moura, and Massih-Reza Amini. "Multitask learning for fine-grained twitter sentiment analysis". In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 2017, pp. 1005–1008.

[KF17] Mathias Kraus and Stefan Feuerriegel. "Decision support from financial disclosures with deep neural networks and transfer learning". In: *arXiv e-prints*, arXiv:1710.03954 (Oct. 2017), arXiv:1710.03954. arXiv: 1710.03954 [cs.CL].

[Dev+18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[HR18] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *arXiv e-prints*, arXiv:1801.06146 (Jan. 2018), arXiv:1801.06146. arXiv: 1801.06146 [cs.CL].

[Pet+18] Matthew E. Peters et al. "Deep contextualized word representations". In: *arXiv e-prints*, arXiv:1802.05365 (Feb. 2018), arXiv:1802.05365. arXiv: 1802.05365 [cs.CL].

[Ara19] Dogu Araci. "Finbert: Financial sentiment analysis with pre-trained language models". In: *arXiv preprint arXiv:1908.10063* (2019).

[Pas+19] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[PCL20a] Yifan Peng, Qingyu Chen, and Zhiyong Lu. "An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining". In: *Proceedings of the 2020 Workshop on Biomedical Natural Language Processing (BioNLP 2020)*. 2020.

[PCL20b] Yifan Peng, Qingyu Chen, and Zhiyong Lu. "An empirical study of multi-task learning on BERT for biomedical text mining". In: *arXiv preprint arXiv:2005.02799* (2020).