Separating Classification from Representation:

Financial Entity-Focused Finetuning with DeBERTa

Keywords: DeBERTa, Targeted Sentiment Analysis, Aspect-Based Sentiment Analysis, Financial NLU

Sean Fuller
Invesco US
Seattle, WA
Sean.Fuller@invesco.com

Abstract

Despite groundbreaking progress across a variety of natural language domains in recent years, research that is directly applicable to investment industry use cases remains scarce. Portfolio managers, traders, and market makers are naturally concerned with specific entities—namely companies—and the degree (i.e., magnitude, not just direction) to which new information might affect the prices of those entities' securities. Thus, sentiment analysis tasks that do not address entity-specific sentiments or that utilize simple positive/negative class labels fall short of the necessary granularity. This paper proposes an entity-focused approach for analyzing fine-grained sentiment expressions in financial news headlines. Using the headlines dataset from Task 5 of the 2017 SemEval competition, we utilize DeBERTa to demonstrate the performance improvement provided by disentangled attention for targeted sentiment analysis. Additionally, we introduce an entity-focused finetuning approach and demonstrate the effectiveness of detaching the classification head from DeBERTa's encoder for multi-token sequence classification tasks. We employ these strategies to produce a finetuned DeBERTa model with a detached CNN-BiLSTM classification head that achieves a .880 cosine similarity on the competition test set, representing state-of-the-art performance on the task. We make our code publicly available for further exploration.

1. Introduction

Approximately 80% of stock market trades are executed by algorithms (Economist 2019). Market events in recent years have revealed that the firms deploying these automated systems utilize natural language data to inform trading decisions. During April of 2013, U.S. stocks momentarily lost $140 billion of market value in under a minute when a hacker gained access to the Associated Press Twitter account and falsely reported an attack on the White House (Johnson 2013). This "flash crash" drew regulatory attention to the costs of unreliable media information in the age of automated trading,

but unreliable processing of factual news poses similar risks. This paper examines how recent advances in natural language processing (NLP) and natural language understanding (NLU) can be leveraged to approach this problem and navigate the associated risks.

2. Literature Review

2.1 Pre-BERT

The dominant framework utilized by most top-performing architectures in competitive NLU tasks today is fine-tuning pretrained contextual word embeddings using a transformers- based architecture. Three important results from 2018 laid the theoretical foundations for this approach. First, the ELMo paper (Peters et al. 2018) introduced the concept of contextual word embeddings. By combining the forward and backwards embeddings of a bidirectional language model, the authors produced word embeddings that can dynamically encode each word based on the context in which it appears. This contextual embedding approach resulted in significant improvement in semantic representation relative to the prior status quo of static embedding lookups.

The second influential paper from 2018 was the ULMFiT paper (Howard and Gugger 2018), in which the authors leveraged insights from the field of computer vision to demonstrate the power of inductive fine-tuning in natural language modelling. Prior to the publication of ULMFiT, adapting pretrained word embeddings trained on large general domain corpora to downstream tasks had been unsuccessful. As a result, competitive architectures in NLU were limited to the approach of building different model formulations on top of static distributed word embeddings. Indeed, this was the approach utilized by all top-scoring contestants in the SemEval Task 5 competition. Unlike previous attempts at LM fine-tuning, Howard and Gugger noted that because different layers of a pretrained model capture different types of information (earlier layers learning general information and later layers learn more specific information), fine-tuning is only successful when different learning rates

are used for different layers of the model and later layers are "unfrozen" from their pretrained states prior to earlier layers. They coined this approach discriminative fine-tuning with gradual unfreezing.

In the third influential paper from 2018, Radford et al. (2018) demonstrated the power of utilizing a transformers-based architecture (Viswani et al. 2017) in the "pretrain fine-tune" paradigm demonstrated by Howard and Gugger. While Howard and Gugger had utilized a simple three-layer LSTM, Radford et al. demonstrated that the multiple head self-attention design of transformers enabled parallelized modelling of more complex semantic relationships between non-contiguous words than LSTMs and other RNN-based architectures.

2.2 BERT and its Variants

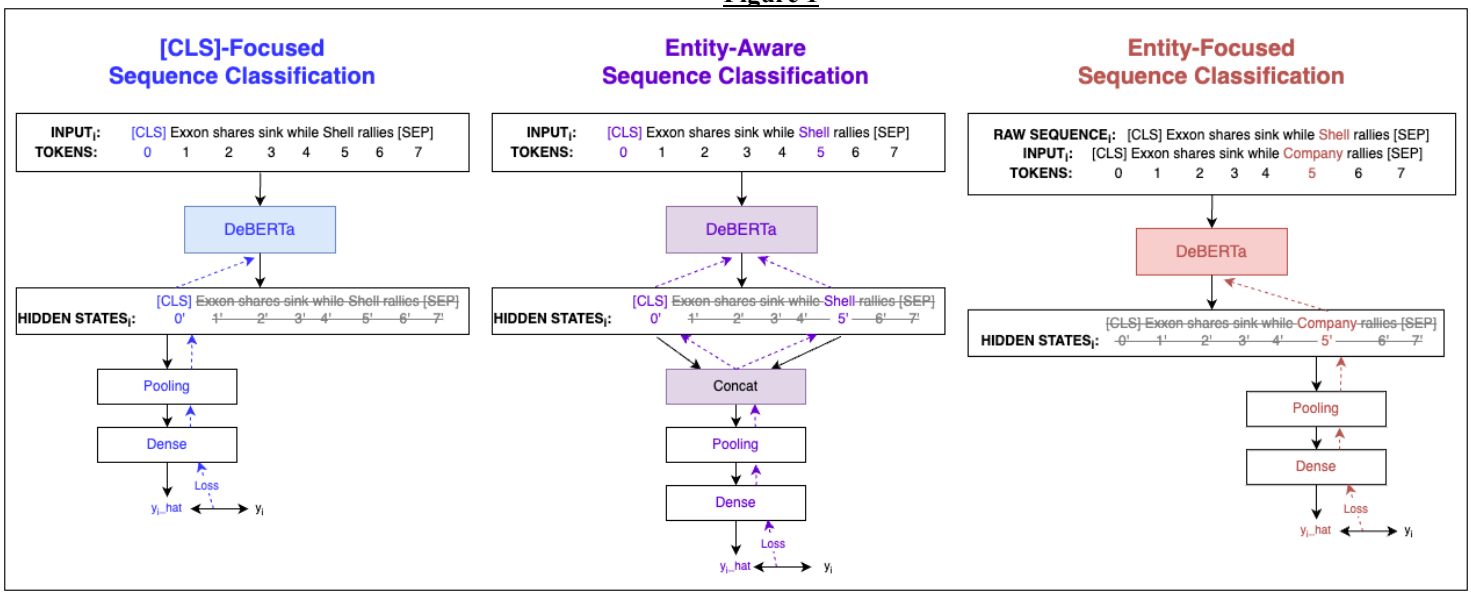
The authors of the BERT paper (Devlin et al. 2019) built on the findings of these three influential papers from 2018 to achieve state-of-the-art performance in many NLU tasks using an encoder-only transformers architecture to produce contextual embeddings within the "pre-train finetune" paradigm. The autoregressive pretraining approach used by Radford et al. (2018) is ideal for generative tasks where the task is next token prediction (like the authors' GPT suite of models). However, noting that downstream tasks that prioritize non-dialogue NLU capabilities (such as summarization and sentiment analysis) would be better modelled by a more aligned pretraining process, Devlin et al. (2019) proposed a bidirectional transformer pretraining approach they coined BERT (Bidirectional Encoder Representations from Transformers). BERT was pretrained using bidirectional masked language modelling (MLM) and next sentence prediction (NSP) to improve on the autoregressive approach by actively targeting contextual representations.

Liu et al. (2019) improved on BERT's results by training on a larger corpus and using multi-token masking, resulting in RoBERTa (Robustly Optimized BERT Pretraining Approach). He et al. (2020) proposed a more material change to BERT in DeBERTa (Decoding-Enhanced BERT with

Disentangled Attention). DeBERTa utilizes separate vectors to encode each word's content and position and maintains separate associated attention weights. DeBERTa holds SOTA status on several NLU benchmarks and was the first model to surpass human performance on the SuperGLUE benchmark. Pretrained weights for DeBERTa are available publicly and are utilized in the Methods section of this paper.

2.3 Related Work

Figure 1



In the BERT paper, Devlin et al. (2019) emphasize the minimalist design of the finetuning step for classification tasks like sentiment analysis. Their finetuning step consists simply of passing the final hidden state of the special [CLS] token to a feed-forward layer followed by a softmax to generate a class prediction. This "traditional" approach is demonstrated in the left-most diagram in Figure 1.

Recent research in targeted aspect-based sentiment analysis (TABSA) and targeted sentiment analysis (TSA) has investigated how BERT can be extended to better model more complex linguistic tasks. Rather than simply predicting a single sentiment metric associated with an input, TABSA and TSA attempt to more precisely model entity-specific sentiment polarities for passages that contain unique expressions of sentiment for multiple referenced entities. The examples from the dataset used

in this paper displayed in Table 1 demonstrate the need for such an approach.

Table 1

id	company	title	sentiment
7	BP	Bilfinger Industrial Services win £100m BP contract extension	0.113
8	Bilfinger Industrial Services	Bilfinger Industrial Services win £100m BP contract extension	0.424

Most influential to this paper is Gao et al. (2019), in which the authors perform target-dependent BERT finetuning by pooling directly on the tokens associated with the target (rather than on the [CLS] token) and feeding the associated hidden state from BERT to the classification head. This approach, which the authors coin Target-Dependent BERT (TD-BERT), is most like the right-most diagram in Figure 1 (though no conversion between the raw sequence and input occurs in their approach). The authors note that this straightforward incorporation of target information shows more stable accuracy improvement than derivative approaches such as feeding the concatenated hidden states of the [CLS] and target tokens to the classification head (as in the center diagram in Figure 1) or applying more complex layers before the classification head (as in the left side of Figure 2). We dig deeper into this observation in our experiments.

Most directly comparable to this paper is Pontes and Benjannet (2021). Utilizing the same dataset used in this paper (SemEval 2017 Task 5), the authors apply a stacked transformer encoder-only layer on top of RoBERTa-generated hidden states enhanced with scores from sentiment dictionaries. The hidden state of the [CLS] token from RoBERTa is then concatenated with the [CLS] token from the secondary transformer network before passing the result to a final feed-forward layer. Compared to the RoBERTa baseline, the authors report a 1.07% improvement in cosine similarity resulting from the secondary transformer stack with sentiment dictionary scoring.

3. Data

The dataset created for Task 5.2 of the 2017 SemEval (Semantic Evaluation) competition uses a fine-grained labelling system, with sentiment scores lying on a continuous numeric scale, on a set of

manually annotated financial news headlines. The downstream task is therefore a natural language regression task rather than one of sentiment classification.

The training data provided for the SemEval 2017 Task 5.2 dataset contains 1,142 news headlines. The test set contains 491 observations. Each observation is labeled with a sentiment score from -1 (most negative) to 1 (most positive). Additionally, a "target entity" is provided for which the sentiment score is to be applied to.

Our primary preprocessing step is to identify and remove recurring headline prefixes. An example is provided in Table 2. This step was not performed when comparing baselines, as indicated by the "Data Cleaned" column in Table 4.

Table 2

id	company	title	cleaned_title
77	AstraZeneca	FTSE 100 movers: Standard Chartered lifted while AstraZeneca sinks	Standard Chartered lifted while AstraZeneca sinks
310	Ashtead	FTSE 100 movers: Ashtead jumps on strong interims; Glencore, BP in ...	Ashtead jumps on strong interims; Glencore, BP in ...
405	ICE	FTSE 100 movers: LSE surges as ICE says mulling offer; Ashtead and Barclays tank	LSE surges as ICE says mulling offer; Ashtead and Barclays tank
498	Barclays	FTSE 100 movers: LSE surges as ICE says mulling offer; Ashtead and Barclays tank	LSE surges as ICE says mulling offer; Ashtead and Barclays tank

The universe of stock tickers is finite, and the SemEval Task 5 training dataset has many repeats of the same target companies. To avoid overfitting to observations of specific company names in the training set, we therefore replace each target company name with the word "Company" for all headlines in the dataset, as shown in Table 3 as well as in the right-most diagrams in Figure 1 and Figure 2. This is indicated by the "Entity Replaced" column in Table 4.

Table 3

id	company	title	entity_replaced_title
4	Glencore	Glencore to refinance its short-term debt early, shares rise	Company to refinance its short-term debt early, shares rise
5	Ryanair	EasyJet attracts more passengers in June but still lags Ryanair	EasyJet attracts more passengers in June but still lags Company

4. Hyperparameters and Methodology

Each architecture in Table 4 was trained on a Huber loss using the Adam optimizer with a learning rate of 2e-5 for 30 epochs with early stopping and a patience of 6. Cosine similarity score is

used as the final evaluation metric, as was the case in the original competition.

To avoid overfitting to the test set, models were not evaluated on the ground truth test set until all models were built and evaluated based on validation set performance. During experimentation, each model was fit five times. The best and median cosine similarities were recorded, and the model producing the best run was saved for each architecture.

For the purposes of comparison to other works, we report the test set performance for each architecture, which was measured during the final step of experimentation. The ranking of our models' performances was identical between validation set and test set evaluation.

5. Experiments and Results

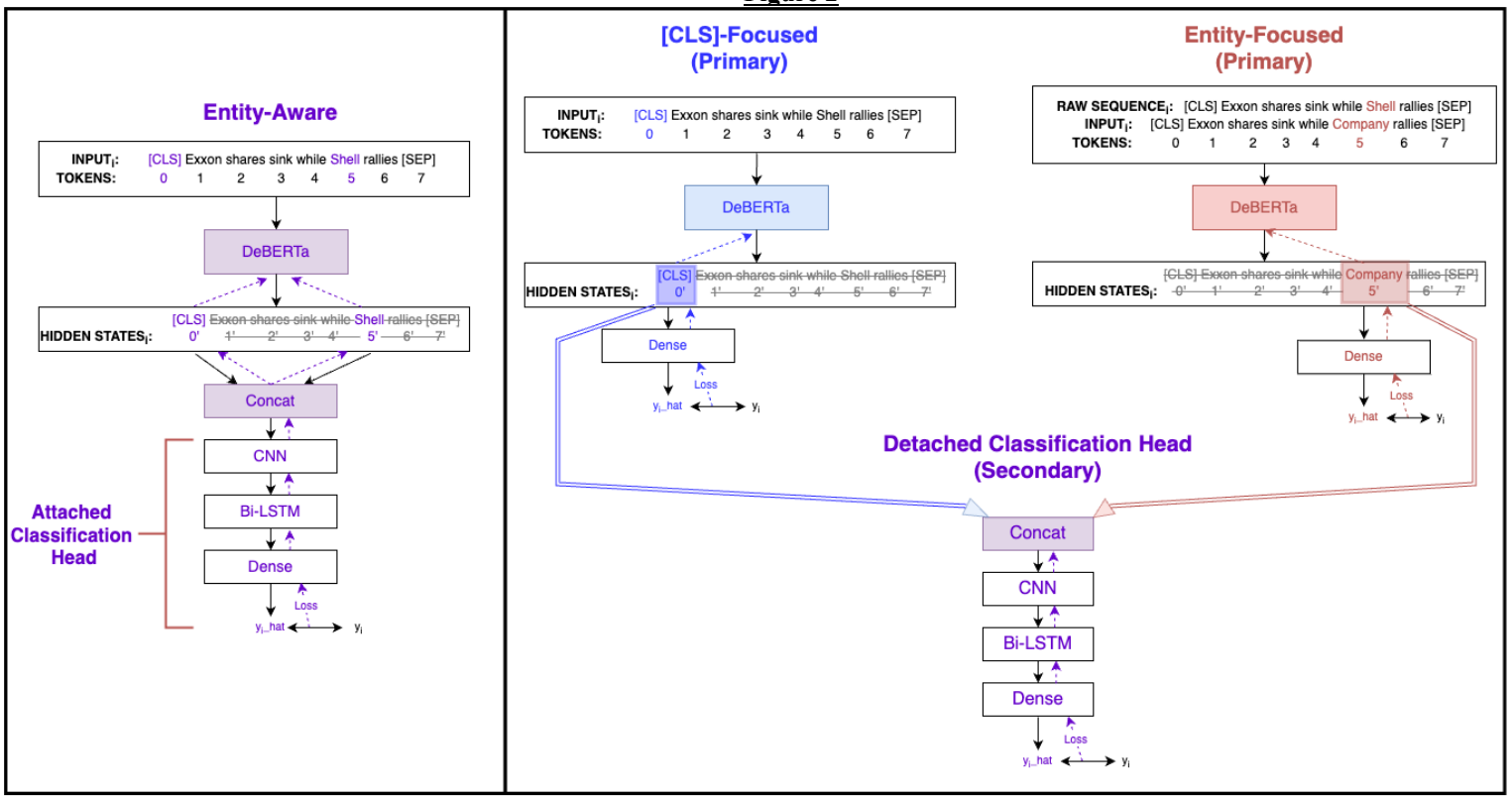
Table 4

Model Groups			Models	Cosine Similarity	Ours	Data Cleaned	Entity Replaced	Pooling Strategy		Classification Head	
								[CLS]	Entity	Attached	Detached
Previous Works	A	1	Mansar et al. (2017)	0.745							
		2	Kar et al. (2017)	0.744							
		3	Rotim et al. (2017)	0.733							
		4	Pontes et al. (2021)	0.848							
Baselines	B	1	BERT	0.793	✓			✓		✓	
		2	FinBERT	0.807	✓			✓		✓	
		3	RoBERTa	0.822	✓			✓		✓	
		4	DeBERTa	0.842	✓			✓		✓	
Token Pooling Strategies	C	1	DeBERTa-[CLS]	0.857	✓	✓		✓		✓	
		2	DeBERTa-[CLS]-Entity-Replaced	0.854	✓	✓	✓	✓		✓	
		3	DeBERTa-Entity	0.874	✓	✓	✓		✓	✓	
Attached CLF Head	D	1	DeBERTa-Attached-FFN	0.866	✓	✓	✓	✓	✓	✓	
		2	DeBERTa-Attached-CNN-BiLSTM	0.849	✓	✓	✓	✓	✓	✓	
Detached CLF Head	E	1	DeBERTa-Detached-FFN	0.879	✓	✓	✓	✓	✓		✓
		2	**DeBERTa-Detached-CNN-BiLSTM**	**0.880**	✓	✓	✓	✓	✓		✓
		3	DeBERTa-Detached-SVR	0.876	✓	✓	✓	✓	✓		✓

We perform four groups of experiments, as delineated in Model Groups B, C, D, and E in Table 4. Group A contains results from the top 2017 competition competitors as well as Pontes and Benjannet (2021) for reference. Group B consists of our comparison between 4 different foundation models: BERT, FinBERT, RoBERTa, and DeBERTa. Noting the performance improvement

provided by disentangled attention, we utilize DeBERTa for the remainder of our experiments.

Figure 2



Group C – Token Pooling Strategies

- **DeBERTa-[CLS]**: This model represents the traditional [CLS]-focused pooling approach presented in the BERT, RoBERTa, and DeBERTa papers. It is identical to the sequence classification implementation provided in popular libraries such as HuggingFace's transformers library. This architecture is represented in blue in Figures 1 and 2. It is the same as Model B4 in Table 4 (the DeBERTa baseline) except that it is fit on the cleaned headlines dataset as explained previously in Section 3.
- **DeBERTa-[CLS]-Entity-Replaced**: This model is identical to **DeBERTa-[CLS]** except that the target entity has been replaced with the word "Company." Classification occurs on the final hidden state of the [CLS] token. Note that performance degrades slightly when the target entity is replaced with the generic "Company" token.

- **DeBERTa-Entity**: This model is similar to the TD-BERT from Gao et al. except that the target entity has been replaced with the generic "Company" token in our model to avoid overfitting to observations containing specific company names in the training data. Classification occurs on the final hidden state of the "Company" token rather than on that of the [CLS] token. This architecture is represented in red in Figures 1 and 2. Like Gao et al., we observe an improvement in performance from this entity-focused pooling approach.

Group D - Attached Classification Head

- **DeBERTa-Attached-FFN**: This model is similar to the TD-BERT-QA-CON from Gao et al. The final hidden states of the [CLS] token and the "Company" token are concatenated before being fed to a fully connected layer (the regression head). Like Gao et al., we notice that this results in a decrease in performance compared to the simpler **DeBERTa-Entity** architecture. This architecture is depicted in purple in Figure 1.
- **DeBERTa-Attached-CNN-BiLSTM**: In this model, the final hidden states of the [CLS] token and the "Company" token are concatenated before being fed to a CNN-BiLSTM classifier with a regression head. This architecture is depicted in purple on the lefthand side of Figure 2. Notably, this model underperformed both of the simpler **DeBERTa-[CLS]** and **DeBERTa-Entity** models**.**

Group E - Detached Classification Head

Interested by the fact that the architectures in Group D utilizing more sophisticated classification/regression heads and multi-token pooling strategies underperformed the simpler models of Group C, we decided to separate the feature extraction and classification portions of the models in Group D. In this group of experiments, we extract the final hidden state of the [CLS] token from **DeBERTa-[CLS]** and the final hidden state of the "Company" token from **DeBERTa-Entity**. We then fit the same

classification architectures from Group D as secondary networks rather than a single unified architecture.

- **DeBERTa-Detached-FFN**: In this version of the model with the regression head separated from the DeBERTa finetuning layers, we observe an improvement in performance over both of the simpler **DeBERTa-[CLS]** and **DeBERTa-Entity** models.
- **DeBERTa-Detached-CNN-BiLSTM**: This model is depicted on the lefthand side of Figure 2. Again, unlike the attached version of this model, we observe an improvement in performance over both of the simpler **DeBERTa-[CLS]** and **DeBERTa-Entity** models. This is our best performing model on both the validation and test sets. This architecture achieves a 2.68% performance improvement over the baseline **DeBERTa-[CLS]** model.
- **DeBERTa-Detached-SVR**: With this model, we demonstrate that even classical machine learning techniques (a support vector regressor in this case) are capable of outperforming deep architectures when the classification head is constructed as a secondary model whose loss does not impact the feature representation during finetuning of the underlying BERT-based model.

6. Analysis and Conclusions

We observe from our experiments that entity-focused finetuning of DeBERTa results in superior performance than the traditional [CLS]-focused approach. This aligns with the results observed in Gao et al. (2019). Additionally, we show that entity-focused finetuning combined with a general "Company" masking strategy is effective for simultaneously improving performance and avoiding company-specific overfitting for financial domain applications.

Our main contribution to the literature is our investigation of the mechanics of the classification head for BERT-based sentiment analysis tasks that are focused on a specific mentioned entity. This problem type is particularly germane to financial domain NLU tasks. We expand on the work of Gao et

al. by observing that separating finetuning from classification enables more performant and stable modelling of entity-focused sentiment by successfully combining sentence-level information (in the [CLS] token) with entity-specific information (in the "Company" token). Our key insight is that when finetuning BERT-based models for sequence classification tasks, the loss is backpropagated through the entire network to update the weights of both the BERT-based model and the final classification layer. The loss from the classification head affects the finetuning process of the underlying BERT-based model, including how the model produces embeddings and converts them to hidden states. We observe that using a single unified network that concatenates the final hidden states of the [CLS] and the entity token prior to the classification head (as in the left side of Figure 2 or in TD-BERT-QA-CON from Gao et al.) results in a finetuned DeBERTa model that underperforms an equivalent two-part network comprised of DeBERTa models that have been separately finetuned to represent sentence-level and entity-level sentiment in isolation (as in the right side of Figure 2 and all models in Group E of our experiments). The intuition behind this is that the classification head of the latter construct is able to learn from each source of information without impacting how their representations are produced by DeBERTa, which we observe to be the superior way of approach classification.

7. Directions for Future Work

Our primary focus has been to show that the separation of classification from representation allows for more effective modelling of entity-specific sentiment expressions in the financial domain. Our exploration of classification architectures has therefore been more expository than exhaustive. We provide links to our code in Appendix A in the hopes that our work encourages further exploration into enhancing classification and regression architectures on top of BERT-based models.

References

Economist. 2019. "The stock market is now run by computers, algorithms and passive managers." The Economist. Oct 5 2019.

Gao, Zhengjie, Ao Feng, Xinyu Song, and Xi Wu. 2019. "Target-Dependent Sentiment Classification with BERT." *IEEE Access*, Vol. 7. https://ieeexplore.ieee.org/abstract/document/8864964.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention." Preprint, submitted June 5, 2020. https://arxiv.org/abs/2006.03654

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-Tuning for Text Classification." Preprint, submitted January 18, 2018. https://arxiv.org/abs/1801.06146

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT*, 1. https://aclanthology.org/N19-1423

Johnson, Steven. C. 2013. "Analysis: False white house tweet exposes instant trading dangers." *Reuters*. April 23, 2013.

Kar, Sudipta, Suraj Maharjan, and Thamar Solorio. 2017. "RiTUAL-UH at SemEval-2017 Task 5: Sentiment Analysis on Financial Data Using Neural Networks." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. https://aclanthology.org/S17-2150

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." Preprint, submitted July 26, 2019. https://arxiv.org/abs/1907.11692

Mansar, Youness, Lorenzo Gatti, Sira Ferradans, Marco Guerini, and Jacopo Staiano. 2017. "Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment Towards Brands from Financial News Headlines." Preprint, submitted April 4, 2017. https://arxiv.org/abs/1704.00939

Pontes, Elvys Linhares, and Mohamed Benjannet. 2021. "Contextual Sentence Analysis for the Sentiment Prediction on Financial Data." *IEEE International Conference on Big Data (Big Data).* https://arxiv.org/pdf/2112.13790.pdf

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." OpenAI. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Rotim, Leon, Martin Tutek, and Jan Šnajder. 2017. "TakeLab at SemEval-2017 Task 5: Linear Aggregation of Word Embeddings for Fine-Grained Sentiment Analysis of Financial News." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. https://aclanthology.org/S17-2148/

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." Preprint, submitted June 12, 2017. https://arxiv.org/abs/1706.03762

Appendix A

For baseline comparisons (Group B in Table 4), please see https://github.com/sfuller14/MSDS-453- Project/blob/master/Fine_Grained_Financial_Sentiment_Regression_with_BERT.ipynb.

For construction and evaluation of the model architectures included in Groups C, D, and E in Table 4, please see https://github.com/sfuller14/DeBERTa_Entity-Focused_Fine-Tuning/blob/master/DeBERTa_Entity-Focused_Fine-Tuning.ipynb. Test set evaluation results are included at the bottom.