Discovering material information using hierarchical Reformer model on financial regulatory filings

François Mercier
Autorité des Marchés Financiers / Mila
Québec, Canada
francois.mercier@lautorite.qc.ca

Makesh Narsimhan
Autorité des Marchés Financiers / Mila
Québec, Canada
MakeshSreedhar.Narsimhan@lautorite.qc.ca

ABSTRACT
Most applications of machine learning for finance are related to forecasting tasks for investment decisions. Instead, we aim to promote a better understanding of financial markets with machine learning techniques. Leveraging the tremendous progress in deep learning models for natural language processing, we construct a hierarchical Reformer ([15]) model capable of processing a large document level dataset, SEDAR, from Canadian financial regulatory filings. Using this model, we show that it is possible to predict trade volume changes using regulatory filings. We adapt the pretraining task of HiBERT ([36]) to obtain good sentence level representations using a large unlabelled document dataset. Finetuning the model to successfully predict trade volume changes indicates that the model captures a view from financial markets and processing regulatory filings is beneficial. Analyzing the attention patterns of our model reveals that it is able to detect some indications of material information without explicit training, which is highly relevant for investors and also for the market surveillance mandate of financial regulators.

KEYWORDS
NLP, Transfer Learning, Financial Reports, Self-Supervised Learning, Material Information detection

1 INTRODUCTION
The entire financial services industry is based on the trust investors have in the system. Preserving this trust is a core mandate from all regulators across the world. Machine learning techniques have the potential to improve several aspects in this industry, including trust. However, most interest so far has been on forecasting tasks for investment decisions.

A better understanding of the markets is useful for everyone in the system. For investors, this translates to a more accurate analysis of the risk-return trade-off. For regulators, it helps to focus their efforts on detecting potential market manipulations and ensuring trust in the overall market.

In this work, we focus on providing answers to the following research questions:
(1) Can we predict market events using only a large document-level dataset from regulatory filings?
(2) If yes, can we leverage potential predictive power to get insights from markets about filings?

The main contributions of our work include:

• Using a hierarchical model, inspired by Hibert (Zhang et al. [35]) and leveraging the efficient Reformer model (Kitaev et al. [15]) for obtaining good representations long sequences and using these obtained representations successfully for a surrogate downstream task, predicting direction of trade volume changes from potentially long documents
• Providing a qualitative assessment of predictions and using attention patterns to better understand the market point of view on focused documents and sentences.

2 PROBLEM DEFINITION

2.1 Background
"Material change", according to securities legislation, is defined as “a change in the business, operations or capital of the issuer that would reasonably be expected to have a significant effect on the market price or value of any of the securities of the issuer and includes a decision to implement such a change made by the board of directors of the issuer by senior management of the issuer who believe that confirmation of the decision by the board of directors is probable”. Regulators require issuers to disclose immediately these ‘material changes’ in order to ensure a level playing field for all investors, and therefore, to ensure trust in financial markets.

2.2 Formal definition
In order to better understand potential material changes, we define the following surrogate downstream task: using documents, publicly available for investors, we try to predict the trade volume movement within a 1 business day time horizon from the release date \( D \) for the stock associated with this document. Let \( V_D \) the daily volume for a specific stock for the day \( D \).

\[
\text{daily volume change} = V_{D+1} - V_D
\]

\[
Y_D = \begin{cases} 
1 & \text{if daily volume change } \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

(1)

The model for this surrogate downstream task is designed to extract focus information at sentences level (see section 4). The sentences receiving the most focus are then selected to propose an extractive summary of these documents, ideally with potential further information about material information.
3 RELATED WORK

3.1 Market signals and market efficiency

Finance is the management of money by companies, organisations, or governments, as defined by Wikipedia\(^1\). It relates to spending and investment of capital, which are decisions made from opinions. Decision processes may be the result of human opinions directly, with manual trade executions, or indirectly, thanks to algorithmic executions using human prior knowledge. With this insight, from a market perspective, indicators, resulting from market activities, are the equilibrium from different investor opinions. These indicators are commonly referred to as market signals. Market signals can be trade prices, prices volatility, limit order book information, or trade volumes.

The main challenge with using market signals, which can be used as direct inputs for investment decisions, is related to the Efficient Market Hypothesis (EMH), Malkiel and Fama\(^2\). The EMH suggests that no one can beat the market without additional risk, which is described with the ‘No Free Lunch’ principle. While there is still a debate regarding the validity of this hypothesis, Fama\(^3\), Malkiel\(^4\), research suggests that the semi strong form of EMH, where all past and present public information is reflected in market prices, seems to hold for developed markets, at least to some extent, see Rossi\(^5\) for literature review. Thus, the consensus is that EMH appears to be a good approximation of market behaviours and one of the main reasons for the difficulty in discovering and using market signals.

In order to tackle this issue, several strategies are often used like incorporating domain knowledge, such as feature engineering with creation of domain specific lexicon from 10-K filings, Loughran and McDonald\(^6\), capturing specific events with open information extraction, Ding et al.\(^7\), and also exploiting new dataset, referred as alternative data, such as Twitter to capture investors sentiment, Smailović et al.\(^8\).

Due to the strong interest of exploiting market signals for investment decisions, most of research focused on price forecasting or price movement direction, (Chong et al.\(^9\), Bartram et al.\(^10\)). Some exceptions can be found for alternative market signals. For price volatility, Theil et al.\(^11\) suggests SEC 10-k filings have got predictive power. Bordino et al.\(^12\) suggests web traffic has predictive power for trade volumes. However, to the best of our knowledge, none are leveraging these market signals as interpretation from markets point of views.

3.2 NLP

Contextual Word Embeddings. Recently, there has been a shift from using distributional word representations (Mikolov et al.\(^13\), Pennington et al.\(^14\)), which result in a single global representation for each word ignoring their context, to contextual embeddings, where each token is associated with a representation that is a function of the entire input sequence. These context-dependent representations can capture many syntactic and semantic properties of words under diverse linguistic contexts. Previous work (Clark et al.\(^15\), Devlin et al.\(^16\), Joshi et al.\(^17\), Liu et al.\(^18\), Peters et al.\(^19\)) has shown that contextual embeddings pretrained on large-scale unlabelled corpora achieve state-of-the-art performance on a wide range of natural language processing tasks, such as text classification, question answering and text summarization.

Document level embeddings. Hierarchical Bidirectional Encoder Representations from Transformers (Zhang et al.\(^20\)) builds upon BERT (Devlin et al.\(^21\)) and proposes a pretraining scheme for document level embeddings. To obtain the representation of a document, they use two encoders: a sentence encoder to transform each sentence in the document to a vector and a document encoder to learn sentence representations given their surrounding sentences as context. Both the sentence encoder and document encoder are based on the Transformer encoder (Vaswani et al.\(^22\)) nested in a hierarchical fashion. They use a variant of the Masked Language Modeling paradigm using sentences as the basic unit instead of words. i.e. they predict masked out sentences given the context. They show that such a pretraining scheme is highly effective and allows them to achieve SOTA results on summarization tasks.

Self-Attention Variants. Recently, there has been a lot of interest in breaking the quadratic self attention used in transformer (Beltagy

\(^1\)https://en.wikipedia.org/wiki/Finance

\(^2\)Malkiel, Burton G., and Eugene Fama. “The Efficient Capital Market Hypothesis: A Review of the Theory and Empirical Evidence.” Journal of Finance, vol. 33, no. 5, 1978, pp. 1631-1668.

\(^3\)Fama, Eugene. “Efficient Capital Markets: A Review of Theory and Empirical Work.” Journal of Finance, vol. 25, no. 2, 1970, pp. 383-417.

\(^4\)Malkiel, Burton. A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing. W.H. Freeman, 2003.

\(^5\)Rossi, Filippo. “The Efficient Market Hypothesis: A Review of Theory and Empirical Evidence.” Journal of Finance, vol. 33, no. 5, 1978, pp. 1631-1668.

\(^6\)Loughran, Todd, and James McDonald. “Lexical Sentiment Analysis.” Journal of Accounting, Auditing & Finance, vol. 24, no. 1, 2009, pp. 135-157.

\(^7\)Ding, Kai, et al. “Entity Type Specific Sentiment Analysis.” Journal of Computing Science and Engineering, vol. 11, no. 3, 2017, pp. 68-75.

\(^8\)Smailović, Aleš, et al. “Investor Sentiment Analysis Using Twitter.” Journal of Behavioral Research in Finance, vol. 6, no. 1, 2016, pp. 1-14.

\(^9\)Chong, Y., and J. N. Bartram. “Market Volatility and Sentiment: Evidence from Twitter.” Journal of Financial Economics, vol. 105, no. 2, 2012, pp. 257-272.

\(^10\)Bartram, J. N., et al. “Market Volatility and Sentiment: Evidence from Twitter.” Journal of Financial Economics, vol. 105, no. 2, 2012, pp. 257-272.

\(^11\)Theil, H., et al. “Market Volatility and Sentiment: Evidence from Twitter.” Journal of Financial Economics, vol. 105, no. 2, 2012, pp. 257-272.

\(^12\)Bordino, Daniele, et al. “Market Volatility and Sentiment: Evidence from Twitter.” Journal of Financial Economics, vol. 105, no. 2, 2012, pp. 257-272.

\(^13\)Mikolov, T., et al. “Efficient Estimation of Word Representations in Vector Space.” arXiv preprint arXiv:1301.3781, 2013.

\(^14\)Pennington, J., et al. “Glove: Global Vectors for Word Representation.” arXiv preprint arXiv:1408.5882, 2014.

\(^15\)Clark, Matthew, et al. “A Character-based Encoder for Sequence Understanding.” arXiv preprint arXiv:1611.01578, 2016.

\(^16\)Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805, 2018.

\(^17\)Joshi, Vibhav, et al. “A Character-based Encoder for Sequence Understanding.” arXiv preprint arXiv:1611.01578, 2016.

\(^18\)Liu, Y., et al. “Robust BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805, 2018.

\(^19\)Peters, Matthew, et al. “A Character-based Encoder for Sequence Understanding.” arXiv preprint arXiv:1611.01578, 2016.

\(^20\)Zhang, Z., et al. “A Character-based Encoder for Sequence Understanding.” arXiv preprint arXiv:1611.01578, 2016.

\(^21\)Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805, 2018.

\(^22\)Vaswani, Ashish, et al. “Attention is All You Need.” arXiv preprint arXiv:1706.03762, 2017.
et al. [2], Kitaev et al. [15], Wang et al. [33], Zaheer et al. [34]) with lower time and memory complexities, enabling the processing of larger sequences and giving rise to better models. Simplest methods in this category just employ a sliding window, but in general, most work fits into the following general paradigm: using some other mechanism select a smaller subset of relevant contexts to feed in the transformer and optionally iterate. In this work, we use the Reformer model which introduces the following improvements: (1) using reversible layers to remove the need to store intermediary activations for the backpropagation algorithm; (2) splitting activations inside the feed-forward layers and processing them in chunks; (3) approximating attention computation based on locality-sensitive hashing.

4 MODEL

4.1 Base hierarchical model: Document representation

We use the following notation: $D = (S_1, S_2, ..., S_{|D|})$ for the sequence of sentences in a document, $S_i = (t^1_i, t^2_i, ..., t^{|S_i|}_i)$ for the sequence of subword tokens for the $i$th sentence and $t^j_i$ for the $j$th subword token for the $i$th sentence.

The base model is the module shared by all the different tasks and is composed of two main submodules, both Reformer models (Kitaev et al. [15]):

- The first submodule transforms the sequence of subword embeddings $(t^1_i, t^2_i, ..., t^{|S_i|}_i)$ into a sequence of contextual embeddings at subword level, $(e^1_i, e^2_i, ..., e^{|S_i|}_i)$ for each sentence $S_i$. For later stages, the first embedding is treated as the sentence embedding for the $i$th sentence.

  $$(e^1_i, e^2_i, ..., e^{|S_i|}_i) = \text{ReformerModel}_{\text{word}}(t^1_i, t^2_i, ..., t^{|S_i|}_i)$$

- The second submodule transforms a sequence of sentence embeddings $(e^1_i, e^2_i, ..., e^{|D|}_i)$ into a sequence contextual embeddings at sentence level, $(d_1, d_2, ..., d_{|D|})$.

  $$(d_1, d_2, ..., d_{|D|}) = \text{ReformerModel}_{\text{sentence}}(e^1_i, e^2_i, ..., e^{|D|}_i)$$

To recap, the base model is defined as follow:

$$(d_1, d_2, ..., d_{|D|}) = \text{BaseModel}(D)$$

For a graphical visualization, please see figure 1.

4.2 Hierarchical model for the pretraining task

For the pretraining task, we used Hibert (Zhang et al. [35]) with the adaptation to long documents by using Reformer (Kitaev et al. [15]) instead of Transformer (Vaswani et al. [31]).

Let $M$ represents the set of indices of the masked sentences.

For the masked sentences with indices $i \in M$, and for each subword tokens with indices $j \in [1; |S_i|]$, we replace all original tokens $t^j_i$ by the mask token <MASK>. We then use a base model to get the contextual embeddings $\{d_i| i \in M\}$ at sentence level for these masked sentences.

As the context, represented by $d_i$, is a fixed vector for each subword prediction, this context is injected by adding this sentence embedding as well as $p^j$ position embedding for the $j$th indice for each timestep. We then use a single Reformer layer as follow:

$$x^j = \text{EmbeddingLayer}(t^j_i) + d_i + p^j$$

$$q^j_{i} = x^j_{i} - s$$

$$k^j_{i} = (x^j_{1}, x^j_{2}, ..., x^j_{i-1})$$

$$v^j_{i} = (x^j_{1}, x^j_{2}, ..., x^j_{i-1})$$

$$h^j_{i} = \text{ReformerLayer}(q^j_{i}, k^j_{i}, v^j_{i})$$

From this, we predict the next word given the masked sentence embedding $d_i$ and the previous words $t^{j-1}_i$ with $W_{ff} \in \mathbb{R}^{\text{vocab size} \times \text{hidden size}}$ and $b_{ff} \in \mathbb{R}$:

$$p(t^{j-1}_i | t^j_i, d_i) = \text{softmax}(W_{ff}h^j_{i} + b_{ff})$$

We then compute the cross entropy loss to be minimized for this task and discard the additional Reformer layer for downstream
tasks.

\[ \text{Loss}_{\text{CrossEntropy}}(D) = -\frac{1}{|M|} \sum_{i \in M} \sum_{j=1}^{|S_i|} \log(p(1|^{j-1}_{i}, d_i)) \]

For a graphical visualization, please see figure 2.

4.3 Hierarchical model for the surrogate downstream task

The classification task aims to incorporate knowledge from market signals into the hierarchical model. For this task, for a document \(D\), the label \(Y_D\) is constructed as per equation 1.

On top of the base model, we add a single Reformer layer with global attention mechanism to retrieve a latent representation of the sequence of sentences \((z_1, z_2, ..., z_{|D|})\) for the document \(D\). The attention weights \(a_i\) used by this Reformer layer will be used later to retrieve the focus of the model.

\((z_1, z_2, ..., z_{|D|}) = \text{ReformerLayer}(\text{BaseModel}(D))\)

We then treat the first latent embedding \(Z_1\) for the document \(D\) as the document embedding and use a classification head. This head outputs the predicted probability of increase or decrease of daily volume, used later for the binary cross entropy loss:

\[ \hat{Y_D} = p(\text{daily volume change} \geq 0|D) = \text{sigmoid}(\text{ClassificationHead}(z_1)) \]

\[ \text{Loss}_{\text{BCE}}(D) = -\left(Y_D \ast \log(\hat{Y_D}) + (1 - Y_D) \ast \log(1 - \hat{Y_D})\right) \]

Finally, to encourage sparsity in the attention weights in additional Reformer layers for this task, in order to increase the focus on the most important sentences, we add a \(l_1\) regularization term for the attention weights \(a_i\).

\[ \text{Loss}_{\text{classification}}(D) = \text{Loss}_{\text{BCE}}(D) + \sum_i |a_i| \]

For a graphical visualization, please see figure 3.

5 EXPERIMENTS AND RESULTS

5.1 Dataset

5.1.1 SEDAR. The System for Electronic Document Analysis and Retrieval (SEDAR)\(^2\) is a filing system developed for the Canadian Securities Administrators (CSA), umbrella organization of provincial and territorial regulators in Canada, to: facilitate the electronic filing of securities information as required by CSA; allow for the public dissemination of Canadian securities information collected in the securities filing process; and provide electronic communication between electronic filers, agents and the CSA. It can be viewed as the Canadian equivalent of EDGAR, the filing system managed by the SEC\(^4\).

For the purpose of this work, we collected 3.8M documents in English from 1997 to October 2018. These documents were originally in PDF and have been converted into raw text format. This conversion has resulted in a certain amount of noise into the raw text.

For the pretraining task, we used 2M documents randomly selected from the overall dataset. All selected documents were before 2018 to prevent information leakage.

For the downstream task, we only used News releases and Management, Discussion and Analysis (MD&A) documents for firms part of the TSX S&P 60 index as of 30th July 2020. The selected document types are among the most known ones to broadcast financial information that are not part of standardised accounting metrics (e.g., net income, cash flow from operations, ...). They are therefore highly susceptible to contain new material information. The training set contained 14,241 documents before 2018. The validation and test set are constituted by all documents in 2018, from which a random allocation has been performed to have 50% for validation set and 50% for test set. This results into 612 documents for validation set and 613 documents for test set. 89.5% of documents contain less than 512 sentences and 94% sentences contain less than 128 subwords.

5.1.2 Market data. For the downstream task, we collected daily trade volumes from 2004 from the Bloomberg Terminal for all components from the TSX S&P 60 Index. We then computed the daily change as per equation 1 and joined this dataset with SEDAR using filing dates and ticker codes. Labels are fairly imbalanced with 59%, 63.6% and 57.8% up movements for train, validation and test sets respectively.

5.2 Implementation details

5.2.1 Tokenization. We used BBPE (Sennrich et al. [28]) implementation from HuggingFace\(^3\) with a vocabulary size 8,000. Similar to common practices in natural language processing, for each sentence \(S_i\), <BOS> and <EOS> are added to represent the beginning of sentence and the end of sentence.

5.2.2 Base model. The base model used for following hyperparameters: 8 attention heads, 8 layers (4 for the Reformer model at word level and 4 for the Reformer model at sentence level), intermediary size 2048, maximum sentence length 128 and maximum number of sentences 512. The number of parameters for the base model was 67M. We used the Reformer implementation from HuggingFace\(^6\).

5.2.3 Pretraining. We trained our model with 2M documents for 1 epoch with a learning rate of 2e−4 with a linear learning rate schedule and with a batch size 32 using gradient accumulation. The duration of the training was 20 days using a single GPU (11GB).

5.2.4 Downstream task. After hyperparameter search, we trained our best model for 2 epochs with a learning rate of 3e−6 for models with frozen base model encoder and 2e−5 otherwise, with a \(l_1\) factor 0.1, with a cosine annealing learning rate schedule and with a batch size 32 using gradient accumulation. We used 2 layers MLP for classification head. Each training run lasted less than 1 day using a single GPU (11GB).

5.3 Evaluation methodology

We evaluated our model performance on the downstream task using the test set with ROC-AUC, MCC and F1 metrics. For a better

\(^{3}\)https://www.sedar.com

\(^{4}\)https://www.securities-administrators.ca

\(^{5}\)https://www.sec.gov/

\(^{6}\)Transformers library: https://github.com/huggingface/transformers

\(^{3}\)Tokenizers library: https://github.com/huggingface/tokenizers

\(^{2}\)https://www.sedar.com

\(^{1}\)https://www.securities-administrators.ca

\(^{4}\)https://www.sec.gov/
results, we used bootstrapping to obtain confidence intervals by repeating 100 times the evaluation on 300 observations. Classifications thresholds were calibrated to maximize MCC and F1 on validation set.

5.4 Results and discussion

5.4.1 Predicting market signals from regulatory filings. Our results, summarized in the table 1, confirm the predictive power from only using regulatory filings for trade volume change prediction.

Our model, without pretraining, outperforms significantly "simple" baselines with the exception of F1 score. F1 score doesn’t take into account true negatives, which penalizes our model compared to the majority class prediction model. Despite the former, our model still achieves similar performance statistically. We only keep F1 score for comparison as this metric is commonly reported in related studies.

As a conclusion, our results support the importance of information from regulatory reporting, at least for News releases and MD&As, for investors in their investment decisions.

5.4.2 Pretraining on document dataset. In this experiment, we compare pretrained models with frozen base model with randomly initialized model with frozen base model and a fully trainable model, also randomly initialized.

As indicated by the table 2, the model pretrained with 2M documents outperforms the fully trainable model with statistical significance on ROC-AUC metric. For MCC metric, the model pretrained with 600K documents is statistically similar to the fully trainable model. For the F1 score, all models are statistically similar, making this metric less useful for ranking. Therefore, models ranking depends on the selected metric, but pretrained models outperform the fully trainable model on all metrics, except MCC for which results are statistically similar.

Moreover, we observed that the best performing model, in term of validation loss, on the pretraining task was the one pretrained on 600K documents. This model is also the best on the MCC metric. For other metrics, it still remains statistically similar to the one pretrained on 2M documents. Whereas the common knowledge assumes that further pretraining helps, see Liu et al. [18], we didn’t find it in our case. As model capacity limited by the single GPU memory constraint, our models were smaller in number of parameters than Zhang et al. [35]. Thus, we believe bigger models may provide better results for this task.

Furthermore, an interesting observation is that, we noticed that pretrained models have less variance on training set and validation set losses during training than the one with frozen weight and random initialization. This suggests that initialization from pretraining does help to have a more robust training.

We conclude that the Hibert (Zhang et al. [35]) pretraining task is slightly beneficial for learning good representation from financial text corpus.

5.4.3 Qualitative analysis. For this experiment, we analyzed the best predictions on validation set from our model, meaning the most confident predictions for increase and decrease trade volumes. We reviewed the documents qualitatively and we also looked at the attention weights at the Reformer layer, on top of the base model, which is specialized for the downstream task, see section 4.3. We applied two attention patterns analysis, one using raw attention weights values similar to the study of BERT attention weights from Clark et al. [6] and one using attention vector norms (Kobayashi et al. [16]). For both analyses, results were fairly consistent, even if some orders slightly changed.

Despite our use of $l_1$ regularizer to encourage sparsity, attention patterns are quite broad among attention heads. This indicates that there is no strong focus on a single sentence, but rather a broad focus on all sentences. While the focus appears to contain some noise, we noticed the model tends to focus more on: the second sentence of the document; sentences containing information about the firm (ex: ticker, firm name, web site url); statements related to disclosures of risks (ex: “Forward-looking statements in this document include, but are not limited to, statements relating to our financial performance objectives, vision and strategic goals, and include our President and Chief Executive Officers statements.”); and also sometimes on mentions to non “Generally Accepted Accounting Principles” (GAAP) measures. While focus on firm name information is quite intuitive, the focus on information about non GAAP measures is quite surprising. Indeed, non GAAP measures have received some attentions from literature and regulators for decades as they may mislead investors in some cases, Bradshaw and Sloan [4], Entwistle et al. [10], Marques [23]. The level of noise in the attention pattern remains an open question. Some reasons may be due to the model capacity, our assumption of using only one document to predict market event without further context, or also to the intrinsic evolving nature of markets.

At the document level, the most confident predictions of decrease of trade volumes appear to be on documents containing no clear new information: such as date confirmation for a result announcement; termination of already announced third party share repurchase program; or third party recognition related to “Environmental, Social and Corporate Governance” (ESG). All of these documents contained no accounting or new business development information. For the most confident predictions of increase of trade volumes, documents contain new accounting information, such as increase of net income or debt refinancing. See figures in appendix for examples. The previously mentioned patterns are consistent with the common knowledge for fundamental analysis.

To conclude, attention weights patterns appear to be too noisy to be used as a clear extractive summary for documents with our approach. At the document level, our model predictions are inline with accepted knowledge about the focus of the market, despite no prior knowledge about markets and no strong labels. Thus, without the need of labelling documents, we believe that our model can be relevant for finding documents containing potential material information, which could help investors for their risk management and also regulators for their market surveillance mandate.

6 CONCLUSION

By leveraging recent progress in natural language processing to process document level large dataset for the financial context, we have proposed a new approach to attempt to discover material information. The key ideas are: efficient deep learning models for long sequences, such as Reformer (Kitaev et al. [15]) in our work; a
Table 1: Test set results for volume direction classification with 95% CI. Majority baseline uses the majority class from training set. Random init.: random weights initialization (no pretraining).

| Model                  | ROC-AUC     | MCC        | F1         |
|------------------------|-------------|------------|------------|
| Random baseline        | [49.9%, 51.0%] | [0.1%, 2.1%] | [53.2%, 54.5%] |
| Majority baseline      | [50.0%, 50.0%] | [0.0%, 0.0%] | [72.7%, 73.4%] |
| Ours - random init.    | [56.9%, 57.8%] | [12.8%, 14.5%] | [72.5%, 73.2%] |

Table 2: Test set results for volume direction classification with 95% CI. Random init.: random weights initialization (no pretraining). Frozen: frozen base model.

| Model                  | ROC-AUC     | MCC        | F1         |
|------------------------|-------------|------------|------------|
| Ours - random init.    | [56.9%, 57.8%] | [12.8%, 14.5%] | [72.5%, 73.2%] |
| Ours - frozen+random init. | [55.1%, 56.2%] | [3.5%, 5.6%] | [72.7%, 73.4%] |
| Ours - frozen+pretrained 600K docs | [57.5%, 58.4%] | [13.6%, 15.2%] | [72.2%, 73.0%] |
| Ours - frozen+pretrained 2M docs | [58.0%, 59.0%] | [11.6%, 13.1%] | [72.2%, 73.0%] |

hierarchical model for capturing sentence level and document level contextualized embeddings; and a surrogate downstream task to align market signals, volume prediction in our work, with financial filings text dataset. We also show the benefits of the HiBERT (Zhang et al. [35]) pretraining task to improve the quality of sentence level embeddings by using a large unlabelled financial corpus. Finally, while attention patterns learnt by our model are still noisy, we were able to demonstrate the ability to discover material information without prior knowledge, which is relevant to regulators for their market surveillance mandate. With this work, we hope to encourage research between deep learning and financial communities as benefits could deserve all actors in the financial industry, including regulators, and ultimately users.

REFERENCES

[1] Sölken M Bartram, Jürgen Bronke, and Mehrdad Motahtari. Artificial intelligence in asset management. Centre for Economic Policy Research (CEPR), 2020.
[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.
[3] Iordino, N. Kourtellis, N. Laptev, and Y. Billawala. Stock trade volume prediction with yahoo finance user browsing behavior. In 2014 IEEE 30th International Conference on Data Engineering, pages 1168–1173, 2014.
[4] M. T. Bradshaw and Richard E G Sloan. Gaap versus the street: An empirical assessment of two alternative definitions of earnings. Journal of Accounting Research, 40:41–66, 2000.
[5] Eunsuk Chong, Chulwoo Han, and Frank C Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83:187–205, 2017.
[6] Kevin Clark, Urvasi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT’s attention. In Proceedings of the 2019 ACL Workshop BlackBoxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.
[7] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR, 2020. URL https://openreview.net/pdf?id=r1xMHlBtVb.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.
[9] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 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), pages 1415–1425, 2014.
[10] Gary M Entwistle, Glenn D Felham, and Chima Mbagwu. The voluntary disclosure of pro forma earnings: a us-canada comparison. Journal of International Accounting Research, 4(3), 2005.
[11] Eugene Fama. Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics, 49(3):283–306, 1998. URL https://EconPapers.repec.org/RePEc:eee:jфинeко:v:49:y:1998:i:3:p:283-306.
[12] Eugene F. Fama. Efficient capital markets. I. The Journal of Finance, 46(5):1575–1617, 1991. doi: 10.1111/j.1540-6261.1991.tb04636.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1991.tb04636.x.
[13] WeiWei Jiang. Applications of deep learning in stock market prediction: recent progress. ArXiv, abs/2003.01859, 2020.
[14] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans, 2020.
[15] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020.
[16] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention module is not only a weight: Analyzing transformers with vector norms, 2020.
[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.
[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.
[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
[20] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x.
[21] Burton G. Malkiel. The efficient market hypothesis and its critics. Journal of Economic Perspectives, 17(1):59–82, March 2003. doi: 10.1257/089533003321164958. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164958.
[22] Burton G. Malkiel and Eugene F. Fama. Efficient capital markets: I. The Journal of Finance, 49(3):283–306, 1994. doi: 10.1111/j.1540-6261.1994.tb04636.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1994.tb04636.x.
[23] Ana Marques. Sec interventions and the frequency and usefulness of non-gaap measures. Review of Accounting Studies, 17(1):59–82, March 2003. doi: 10.1257/089533003321164958. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164958.
[24] Ana Marques. Sec interventions and the frequency and usefulness of non-gaap measures. Review of Accounting Studies, 17(1):59–82, March 2003. doi: 10.1257/089533003321164958. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164958.
[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. volume 14, pages 1532–1545, 01 2014. doi: 10.1111/j.1540-6261.2010.01625.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x.
[26] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations.
Discovering material information on financial regulatory filings

In *Proc. of NAACL*, 2018.

[27] Matteo Rossi. The efficient market hypothesis and calendar anomalies: a literature review. *International Journal of Managerial and Financial Accounting*, 7(3-4):285–296, 2015.

[28] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

[29] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidarič. Predictive sentiment analysis of tweets: A stock market application. In Andreas Holzinger and Gabriella Pasi, editors, *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39146-0.

[30] Christoph Kilian Theil, Sanja Stajner, and Heiner Stuckenschmidt. Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 32–37, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3104. URL https://www.aclweb.org/anthology/W18-3104.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[33] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[34] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2020.

[35] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. *CoRR*, abs/1905.06566, 2019. URL http://arxiv.org/abs/1905.06566.

[36] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. *CoRR*, abs/1905.06566, 2019. URL http://arxiv.org/abs/1905.06566.
Figure 4: Examples of documents within most confident predictions of increased volumes with attention focus based on attention vector norms. Darker color indicates stronger focus.
Figure 5: Examples of documents within most confident predictions of decreased volumes with attentions focus based on attention vector norms. Darker color indicates stronger focus.