Learning Target-Specific Representations of Financial News Documents For Cumulative Abnormal Return Prediction

Junwen Duan[†]*, **Yue Zhang**[‡], **Xiao Ding**[†], **Ching-Yun Chang**[‡], **Ting Liu**[†]
[†]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Harbin, China
{jwduan,xding,tliu}@ir.hit.edu.cn
[‡]Singapore University of Technology and Design
yue_zhang@sutd.edu.sg,chang.frannie@gmail.com

Abstract

Texts from the Internet serve as important data sources for financial market modeling. Early statistical approaches rely on manually defined features to capture lexical, sentiment and event information, which suffers from feature sparsity. Recent work has considered learning dense representations for news titles and abstracts. Compared to news titles, full documents can contain more potentially helpful information, but also noise compared to events and sentences, which has been less investigated in previous work. To fill this gap, we propose a novel target-specific abstract-guided news document representation model. The model uses a target-sensitive representation of the news abstract to weigh sentences in the news content, so as to select and combine the most informative sentences for market modeling. Results show that document representations can give better performance for estimating cumulative abnormal returns of companies when compared to titles and abstracts. Our model is especially effective when it used to combine information from multiple document sources compared to the sentence-level baselines.

1 Introduction

Texts from the Internet was shown to be statistically correlated with stock market trends (Antweiler and Frank, 2004). Natural language processing (NLP) techniques have been applied to extract information from company filings (Lee et al., 2014), financial news articles (Xie et al., 2013) and social media texts (Bollen et al., 2011) in order to gain understandings of financial markets. In particular, traditional methods exploited statistical signals, such as lexical and syntactic features (Wang and Hua, 2014; Schumaker and Chen, 2009), sentiment (Bollen et al., 2011) and event structures (Ding et al., 2014; Ding et al., 2015; Ding et al., 2016) from these text sources, which suffers from feature sparsity problems.

With the recent trends in deep learning for NLP, neural networks have also been leveraged to learn dense representations for text elements, which can address the sparsity of discrete features in statistical models. Such representations can implicitly represent sentiment, event and factual information, which can be extremely challenging for sparse indicator features. In particular, Ding et al. (2015) show that deep learning representations of event structures yield better accuracies for stock market prediction compared to discrete event features. Chang et al. (2016), Duan et al. (2018) use neural networks to directly learn representations of news abstracts, showing that it is effective for predicting the cumulative abnormal returns of public companies.

One limitation of Ding et al. (2015) and Chang et al. (2016), however, is that these methods only model news titles and abstract texts, which are typically single sentences. Ding et al. (2015) show that a model that uses only news content gives inferior results compared to one trained using news titles only, and that adding news content information to a title-driven model does not significantly improve the results. Intuitively, news content can contain richer information that is not directly relevant to the title message, or the most important event, and hence can lead to noise in predictive modeling. On the other hand, news

* This work was done while the first author was visiting Singapore University of Technology and Design. Yue Zhang is the corresponding author.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2823–2833
Santa Fe, New Mexico, USA, August 20-26, 2018.

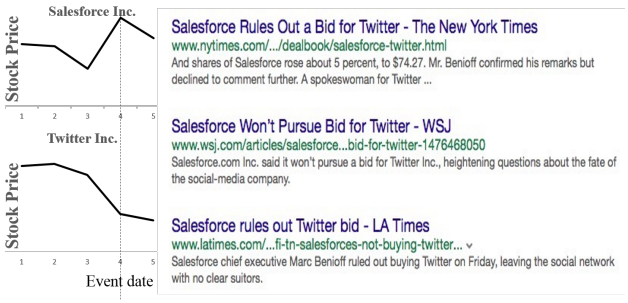

Figure 1: Event "Salesforce rules out Twitter bid" showing different impacts to differnt firms. Twitter fell 5.12%, while Salesforce rose 5.15% on the event day.

content can also contain useful information for making informed decisions. For example, given the news abstract passage *"Amicus Inc (arcs.o), a provider of health-care IT services, said it agreed to be bought by an affiliate of private firm them bravo llc for $217 million in an all-crash deal."*, it would be difficult to tell whether the acquisition is beneficial to investors. However, a sentence in the news content states that *"the deal offers Amicus Inc shareholders $5.35 for each share, a premium of 21 percent over the stock's close on December 24 on NASDAQ"*, which explicitly indicates a positive return.

We aim to exploit such useful information from the news content for making more informed decisions in stock market prediction. A main challenge is how to automatically identify the most useful parts of the news content, while disregarding noise, which prevents naive utilization of news contexts (aka Ding et al. (2015)). Another challenge, as Chang et al. (2016) suggest, is that information must be selected with regard to a certain stock of interest. As shown in Figure 1, the same event "*Salesforce rules out Twitter bid*" can lead to different influences on different stocks, with Salesforce benefiting from it yet Twitter suffering from it.

To address the above challenges, we build a neural model that selects and represents relevant sentences from a full news document with respect to a specific firm of interest. In particular, we leverage conditional encoding (Rocktäschel et al., 2015) to encode information of a given stock into the dense representation of the news abstract. Using this target-specific news abstract representation, we apply neural attention (Bahdanau et al., 2014) over each sentence in the news content to automatically learn its relative importance with regard to the target company. The attention weights are learned automatically towards a final predictive goal. The model is full data-driven, which do not rely on an external syntactic parser as the first step to obtain its target-specific linguistic structures.

In addition to Chang et al. (2016) model, which uses only abstract information, we also compare with several state-of-the-art baselines for learning document representations, such as paragraph vector (Le and Mikolov, 2014) and hierarchical attention network (Yang et al., 2016), giving the best reported performances. The advantage of our approach over sentence-level baseline is especially obvious when it is used to combine information from multiple news document sources. In addition, a case study shows that our model can select the sentences that most intuitively help predict stock returns from a full news document. Our contributions can be summarized as follows:

- We propose a target-specific document representation model, which leverages the abstracts as evidences to select informative sentences from the documents while disregarding noise.
- We are the first, to our knowledge, to build a neural model that can effectively leverage full news document for stock market prediction.

  Resources of this work can be found at http://github.com/sudy/coling2018

2 Problem Definition

Cumulative Abnormal Return The task that we attack in this paper is *Cumulative Abnormal Return Prediction (**CAR**)*. Formally, the abnormal return AR_{jt} of a firm j on a date t is the difference between its actual return R_{jt} and the expected return $\hat{\mathrm{R}}_{jt}$, $\mathrm{AR}_{jt} = \mathrm{R}_{jt} - \hat{\mathrm{R}}_{jt}$. The expected return $\hat{\mathrm{R}}_{jt}$ can be

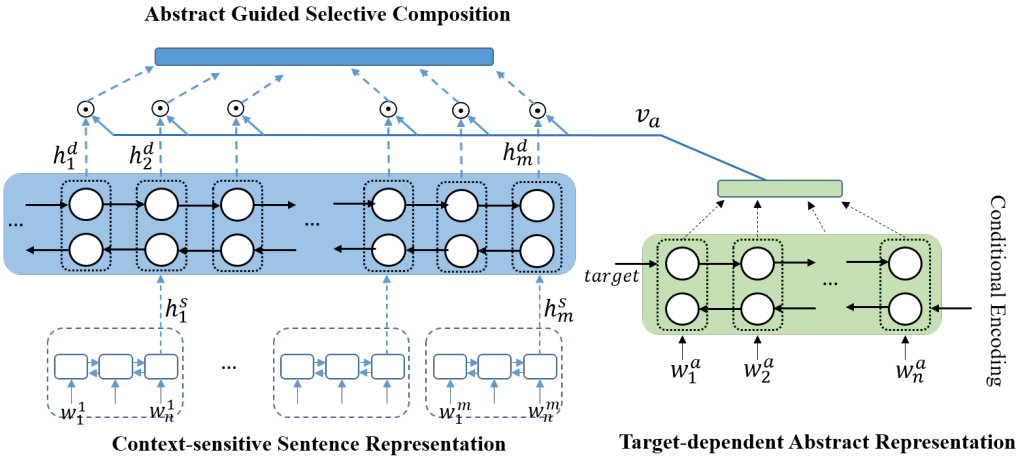

Figure 2: The architecture of proposed method

estimated by an asset price model based on historical prices, or approximated by the market return in a short-term event window (Kothari and Warner, 2004). The cumulative abnormal return CAR_j of the firm j in an n-day time window is calculated by summing up the daily abnormal returns in the period (Eq 1). In this paper, we adopt the commonly used three-day window (-1,0,1), which we denote as CAR_3 and day 0 is the day when the current news documents are released.

$$\text{CAR}_j = \sum_{t=1}^{n} \text{AR}_{jt} \tag{1}$$

Cumulative Abnormal Return Prediction The ultimate goal is to build a model that predicts the cumulative abnormal return CAR_3 of a public company based on a set of related news documents released on day 0. Given a document D and associated firm f, we learn a representation d_f for D that is specified to firm f based on our proposed approach (introduced in Section 3). With the firm-specific document representation d_f, we predict the CAR_3 direction $y \in \{-1, 1\}$ using a parameterized softmax function (Eq 2).

$$p(y|d_f) = \text{softmax}(W \cdot d_f + b), y \in \{-1, 1\}, \tag{2}$$

where $p(y = -1|d_f)$ and $p(y = 1|d_f)$ indicate the probability of CAR_3 being negative and positive, respectively. In the next section, we describe our method for learning d_f.

3 Model

The challenges for learning a target-sensitive document representation are two-fold. On the one hand, we must encode firm-specific information into the dense document representations so as to make them different across targets. On the other hand, we must identify the most informative sentences while disregarding noise for the prediction. To remedy this, we propose to first learn a target-specific representation for the abstract, which is most relevant to the market movement among all the sentences and then leverage the abstract to guide the sentence selection. The architecture of our proposed approach is illustrated in Figure 2, which consists of three key modules. We give details of each module in this section.

3.1 Target-dependent Abstract Representation

As the first step, we learn a target-dependent representation for the abstract of a news document by encoding target information into it. We use a bidirectional Long-short memory network as the basic model. In order to allow information of a target company to influence the semantic representation, we apply conditional encoding (Rocktäschel et al., 2015), using an embedding vector of the target firm of concern $e^t(c)$ as the initial state vector for the sentence-level Bi-LSTMs. We average the hidden states of each word in the sentence to obtain the target-dependent news abstract representation v_a. The vector

$e^t(c)$ for the company of interest is initialized by averaging the words of its constituents and are fine-tuned during training.

3.2 Context-sensitive Sentence Representation

To preserve the semantic structures of documents and make the sentence representations aware of their contexts, we leverage a hierarchical structure (Li et al., 2015) to encode sentences (the abstracts are not considered) in a document. A sentence-level LSTM is first used to encode words into hidden state vectors, and then a document-level LSTM is applied to encode sentences into hidden state vectors.

At the sentence level, given a sentence $\{w_1, w_2, \cdots, w_n\}$, we obtain their embedding forms $\{\vec{e}(w_1), \vec{e}(w_2), \cdots, \vec{e}(w_n)\}$ via a lookup table. A Bi-LSTM is used to capture sentence-level context from $\{\vec{e}(w_1), \vec{e}(w_2), \cdots, \vec{e}(w_n)\}$, yielding two sequences of hidden states $\{\overrightarrow{h_1^w}, \overrightarrow{h_2^w}, \cdots, \overrightarrow{h_n^w}\}$ and $\{\overleftarrow{h_1^w}, \overleftarrow{h_2^w}, \cdots, \overleftarrow{h_n^w}\}$, respectively. We average $\overleftarrow{h_i^w}$ and $\overrightarrow{h_i^w}$ into a single vector h_i^w for the word w_i, and use average pooling over the sentence to obtain a single sentence embedding vector h^s.

In the document level, given the sentence embedding $\{h_1^s, h_2^s, \cdots, h_m^s\}$ for sentences $\{s_1, s_2, \cdots, s_m\}$ in a document D, respectively, the same Bi-LSTM structure (with a different set of model parameters) is applied to obtain hidden states $\{\overrightarrow{h_1^d}, \overrightarrow{h_2^d}, \cdots, \overrightarrow{h_m^d}\}$ and $\{\overleftarrow{h_1^d}, \overleftarrow{h_2^d}, \cdots, \overleftarrow{h_m^d}\}$, respectively. For each sentence s_i, the forward and backward hidden vectors are averaged to give a single hidden state embedding h_i^d. h_i^d contains both internal information from the sentence by semantic composition of words, and document level context by bi-directional recurrent composition.

We use this vector form of each sentence as their representation for abstract and sentence composition. We apply the same conditional encoding to sentence-level and document-level Bi-LSTMs in context-sensitive sentence representation.

3.3 Abstract Guided Selective Composition

As mentioned in Section 1, some sentences can give background information that can support decision making, which is too lengthy to include in the abstract. To address the challenge of informative sentence selection, we consider the target-specific representation obtained as well as the sentence relevances to the abstract in our model.

To compose context-sensitive sentence representations into a final document representation, we use the attention mechanism (Bahdanau et al., 2014), taking the embedding of the abstract to guide calculation of attention weights. Another benefit of using the attention mechanism is the prediction is made interpretable, since interpretability is crucial in financial prediction, as investors have to verify the underlying basis for the decisions.

Formally, given the target-specific representation of the abstract v_a from Section 3.1 and the context-sensitive sentence representations $\{h_1^d, h_2^d, \cdots, h_m^d\}$ from Section 3.2, we concatenate v_a and h_i^d and feed it to a single-layer, feed-forward neural network (Eq 3) to get the score for each sentence h_i^d in the document,

$$u_i = v^\top \tanh(W_a[v_a, d_i] + b), \tag{3}$$

where W_a is the weight matrix, b is the bias, v is a projection vector that maps the output to a scalar and u_i is the weight score showing how much attention should be put on current sentence s_i in the composition of the whole document. The attention score α_i are computed from u_i by Eq 4.

$$\alpha_i = \frac{\exp(u_i)}{\sum_i \exp(u_i)}, \tag{4}$$

$$d = \sum_i \alpha_i h_i^d. \tag{5}$$

Here α_i is the normalized weight score, where $\sum_i \alpha_i = 1$. The final output d is a weighted sum of all the hidden states h_i^d (Eq 5). We concatenate the vector of the abstract v_a and d as the final representation for the document $d' = [v_a, d]$.

	Training	Development	Test
+CAR$_3$	9,674	493	995
-CAR$_3$	9,643	507	1,005

Table 1: Number of CAR$_3$ in the datasets

3.4 Multi-Document Composition

There can be more than one news document mentioning a firm of interest in a certain event window. We adopt a self-attentive neural network (Bahdanau et al., 2014) to assign different weights to the news documents $\{d'_1, d'_2, \cdots, d'_n\}$ (Eq 6) with respect to our prediction goal.

$$u_i = v^\top \tanh(W^{(s)} d'_i + b^{(s)}), \tag{6}$$

where $W^{(s)}$ is a weight matrix and $b^{(s)}$ is a bias, u_i is a scalar showing how much attention should be put on document d_i. The normalization process is the same with Eq 5 but with different parameters. We denote the final representation as $\hat{d}$, and use it as d_f for CAR$_3$ prediction in Section 2.

4 Training

Given a set of training instances T with documents and their corresponding CAR$_3$ as discussed in Section 2, the overall training objective is to minimize the cross-entropy loss with `L2` regularization (Eq 7),

$$\min_{\Theta} - \sum_{T} \sum_{j} t_j \log p(y_j|d_f) + \sum_{\theta \in \Theta} \lambda_\Theta ||\theta_2^2||, \tag{7}$$

where $p(y_j|d_f)$ is the predicted distribution for class j and t_j is the ground-truth distribution. Θ represents all the parameters and the λ_Θ are the regularization parameters.

We pre-train continuous bag-of-word (CBOW) word embeddings with dimension 100 on a collection of Reuters and Bloomberg financial news articles, which is released by Ding et al. (2014). The overall document size is over 400K. The open source toolkit *word2vec*[1] is used to train the word embeddings, which are then fine tuned during the training of the prediction model to help boost the model performance. Out-of-vocabulary (OOV) words are replaced by a `UNK` token. We adopt `Adam` (Kingma and Ba, 2014) as our optimization algorithm, with the initial learning rate being set to 0.0005, `L2` regularization at the strength of 10^{-6}. We use mini-batch size of 32, the model that achieved the best micro-F1 performance on the development set is kept for final test.

5 Experiments

5.1 Data

We collect publicly available financial news articles from Reuters from October 2006 to December 2015. In our preliminary experiments, we find that a news document is more likely to be relevant to a firm only if it is mentioned in the news abstract. Thus, we only include news documents mentioned at least one public listed firm in the U.S. security market. We group the news documents per firm per event date. If the news is released during a trading hour, day 0 is the current day, otherwise day 0 is the next trading day. We compute the expected return $\hat{\mathrm{R}}_{jt}$ by the return of equally-weighted market index including all the stocks on NYSE, Amex, NASDAQ.

We follow Chang et al. (2016) and split the dataset into training, development and test sets, with 1000 instances for development, 2000 instances for testing and the rest for training. The numbers of positive and negative CAR$_3$ in the dataset are listed in Table 1, in which the positive and negative cases are fairly balanced.

[1] https://code.google.com/archive/p/word2vec/

5.2 Evaluation Metrics

Our model is designed for stock recommendation, where companies with high absolute cumulative abnormal return expectations can be given to traders for their information. As a result, it can be useful to have a confidence threshold $beta$ in the model, so that only companies with $p(y = 1|d_f) > \beta$ or $p(y = -1|d_f) > \beta$ are recommended. With increasing β values, the number of recommended stocks is expected to decrease, while the precision to increase. As a result, the model performance is evaluated by the area under the precision-recall curve (AUC). The precision and recall are calculated on both the positive class and negative class. This metric offers a trade-off between precision and recall when varying the prediction confidence threshold. Following Chang et al. (2016), we also evaluate the model on test instances with $|\text{CAR}_3| > 2\%$, namely the event windows with high impacts.

5.3 Baselines

We compare our method with a set of baseline representation learning methods for d_f learning including *target-dependent abstract* representation method of Chang et al. (2016), and state-of-the-art *target-independent document* representations.

Target-dependent News Abstract representation (TGT-CTX-LSTM) Chang et al. (2016) give the current state-of-the-art accuracies for CAR prediction by using the abstract only. They used a syntactic parser to analyze the dependency structure of the abstract (as a sentence), and then transform it to a target-specific dependency tree form. They leverage a tree structured LSTM (Tai et al., 2015) to learn abstract representation according to the target-specific tree form.

Paragraph Vector We take the paragraph vector embedding of Le and Mikolov (2014) as an unsupervised full document representation baseline. This method gives remarkable performances on tasks such as document classification (Yang et al., 2016) and sentiment analysis (Tang et al., 2015).

Target-dependent sentence combination (TD-AVG) This model first learns a context-sensitive representation for all sentences using Bi-LSTM and conditional encoding (Rocktäschel et al., 2015). Average pooling is then applied on the sentence representations to obtain the final document representation. This baseline is equivalent to a conditionally encoded version of the target-specific model of Li et al. (2015), but with a different training objective (i.e., classification instead of auto-encoder). We refer to it as TD-AVG. We use conditional encoding (Rocktäschel et al., 2015) on the sentence-level and document-level Bi-LSTMs.

Target-dependent hierarchical neural network (TD-HAN) This model is similar to TD-AVG except that it adopts an attention model (Bahdanau et al., 2014) over the context-sensitive sentence representations. Intuitively, similar to abstract, some parts of the document are more related to CAR compared to others. Such sentences should be given more weights in composition for document representations. This baseline is adapted from Yang et al. (2016), who applied attention on both the word level and the sentence level in a hierarchical LSTM network for document representation. In the implementation, it is slightly different in not using attention on the word level. We refer to it as TD-HAN.

5.4 Results

Table 2 summarizes the AUCs of both positive and negative classes of different embedding models on the test dataset. As shown in Table 2, our model gives AUCs of 0.65 for both positive and negative classes on the test dataset, outperforming all the baselines. Following Chang et al. (2016), we also make comparisons on cases that give higher CAR, with threshold $|\text{CAR}_3| > 2\%$. This subset covers a total of 1021 cases. Our model gives AUCs of 0.75 and 0.73 on $+\text{CAR}_3$ and $-\text{CAR}_3$, respectively. The improvement is statistically significant at a 1% significance level using T-test. Figure 3 presents the precision-recall curve of *Paragraph Vector*, *TGT-CTX-LSTM* (Chang et al., 2016) and our proposed method. We make more detailed analysis of our approach and relevant models.

	+CAR$_3$	-CAR$_3$	+CAR$_3$ $> 2\%$	-CAR$_3$ $< -2\%$
Paragraph Vector (Le and Mikolov, 2014)	0.51	0.53	0.54	0.55
TGT-CTX-LSTM (Chang et al., 2016)	0.63	0.62	0.70	0.68
TD-AVG (Tang et al., 2015)	0.61	0.61	0.71	0.68
TD-HAN (Yang et al., 2016)	0.63	0.63	0.72	0.71
Our Approach (without target)	0.64	0.63	0.73	0.72
Our Approach	**0.65****	**0.65****	**0.75****	**0.73****

Table 2: Final *AUC* on the test set, ** means that the result is better than the best baseline at 1% significance level.

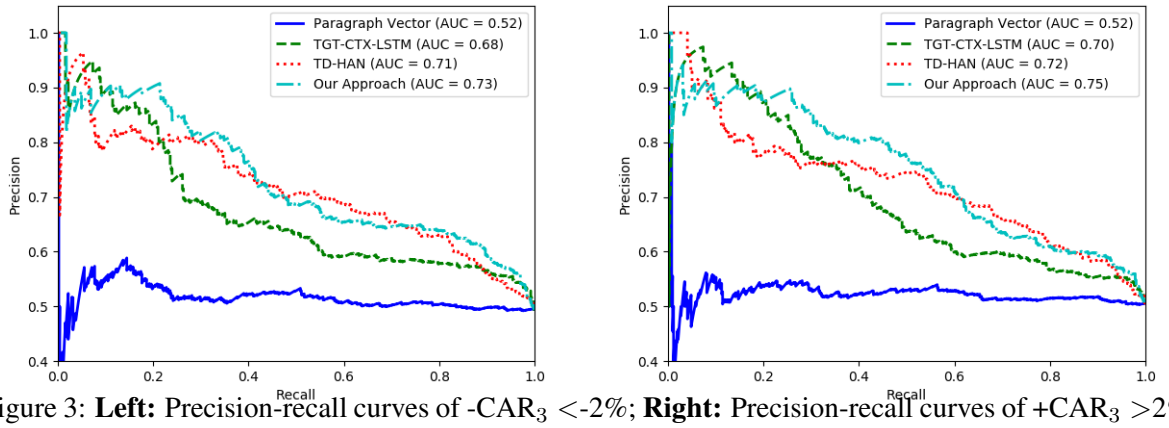

Figure 3: **Left:** Precision-recall curves of -CAR$_3$ $<$-2%; **Right:** Precision-recall curves of +CAR$_3$ $>$2%

The Usefulness of Supervision The performance of *Paragraph Vector* falls far behind other approaches. The primary cause is that its document representation is both target-independent and unsupervised, which neither benefit from the end task nor information from specific targets. This demonstrates the importance of supervisions in capturing the deep semantic meaning of documents for our prediction task.

The Usefulness of Modeling Full Document As shown in Table 2, simple averaging schema TD-AVG gives inferior performances compared to the target-specific abstract representation TGT-CTX-LSTM. This indicates the informativeness of abstract for predicting the stock market movements. However, when the simple average pooling of TD-AVG is replaced by a simple feed-forward attention over the context-sensitive sentence representation, i.e., TD-HAN, the model achieves comparable performance with state-of-the-art model TGT-CTX-LSTM. The results indicate the importance of choosing the right strategy to compose the sentences in document-level compositions, since a document is more sparse and contains noise. It also implies that if the document-level background information is properly modeled, document-level models have obvious advantages over sentence-level model, since short text and simple structure may not fully represent the information conveyed in the document.

Our proposed method, which incorporates firm-specific information as well as the hierarchical structures of news documents, achieves the best performances, showing that target-specific supervisions and fine-grained information from the main contents of news can help gain understanding about the market fluctuations.

The Impact of Target The performances of our model with and without target information are also compared in Table 2. Our model without target information achieves 0.64 and 0.63 AUC on positive and negative classes, respectively, which is comparable with the best baselines. Encoding the target information into the representation did not bring a significant improvement to the model performance. An explanation for the relatively small difference between our method with and without target-specific

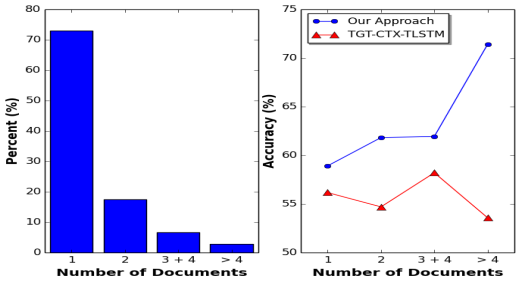

Figure 4: **Left**: The proportion of instances with different number of financial news documents in our test dataset. For example, over 70 percent of instances in our test dataset have information collected from one document in the event window; **Right**: The accuracy of our approach and baseline TGT-CTX-TLSTM against the number of news documents in the event window.

encoding is that there is a relatively low ratio (10.2%) of news abstracts in our dataset mentioning more than one company.

Accuracy with Respect to the Number of Documents As mentioned in the previous section, there can be more than one news documents associated with a firm of interest in a particular event window. More than 25 percent of CAR_3 instances in our dataset have information gathered from more than one documents. Extra documents can provide richer background information about the firm of interest, but may also bring more noise. Figure 4 (Right) shows the accuracies of our model and the baseline TGT-CTX-LSTM with respect to the number of news documents. It is worth noting that our model and TGT-CTX-LSTM use the same attention strategy to combine information from multiple documents. The performance of our approach improves as the number of documents to model grows, showing its effectiveness in utilizing more non-overlapping information sources. Compared to sentence-level baseline TGT-CTX-LSTM, our model can better capture information from multiple document sources.

5.5 Case Study

To illustrate the attention mechanism for weighing sentences in the news documents, Table 3 shows the details of three news documents. The target firm, the actual CAR_3 and the prediction probability of our method and the baseline method TGT-CTX-LSTM are shown in the first rows of each news. The first line of each document is the abstract. The automatically learned weights are shown in the front of each sentence. For case 1, since the most informative parts of the document are not in the abstract, TGT-CTX-LSTM, which only exploits the abstract, yields the incorrect prediction. Our model puts more weights on the last two sentences, which suggest the value of the acquired properties and the voting rights for the target Lennar Corp.. This meets human expectation, since they have shown the potential of the future prospect of the firm.

For case 2, though the ground-truth CAR_3 outperforms the market by a large margin, the baseline TGT-CTX-LSTM shows less confidence for the prediction compared with our method. The abstract gives little knowledge about the purchase event, which make the system that rely on the abstract difficult to make a decision. The second sentence in the news explicitly state the impact of the news to the shareholders, the target firm stock price rose by more than 20 percent. We observed similar patterns in other cases, where the model prefer to put more weights on the sentences which explicitly mention the stock returns. An explanation is that the market follows the trends and it takes time for the market to absorb all information.

For case 3, both methods present similar confidence for the prediction. However, our model puts more weight on the third sentence instead of the abstract. The two sentences can be treated as paraphrases, because they are very close in the meaning, both mentioning future profit expectations of the firm and a quarterly profit decline. It is worth noting that, for documents with a large number of long sentences, our model tends to give evenly-distributed weights for all the sentences in such documents. This is a limitation of our attention model, which we leave for future work. Above all, the result demonstrates that our model is able to automatically capture the most informative parts in the documents.

Our Method	TD-HAN	Case: No. 1; Target: Constellation Brands Inc. ; CAR_3:3.3% ; Our method:0.57; TGT-CTX-LSTM:0.45
0.05	0.07	**Abstract**: builder lennar_corp (len.n) said on friday it formed an investment venture with morgan stanley real estate and sold the venture $ 525 million in properties .
0.07	0.28	the properties acquired by the new entity consist of about 11,000 homesites in 32 communities across the united states , it said .
0.07	0.30	lennar_corp and morgan stanley real estate , an affiliate of morgan stanley & co. (ms.n) , formed the venture to acquire , develop , manage and sell residential real estate .
0.34	0.24	lennar_corp acquired a 20 percent ownership interest and 50 percent voting rights in the venture .
0.47	0.09	as of september 30 , the acquired properties had a net book value of $ 1.3 billion , it said .
Our Method	**TD-HAN**	**Case**: No. 2; **Target**: Amicas Inc.; CAR_3:23.2%; **Our method**:0.77; **TGT-CTX-LSTM**:0.53
0.06	< 0.01	**Abstract**: amicas_inc amcs.o , a provider of healthcare it services , said it agreed to be bought by an affiliate of private firm thoma bravo llc for $ 217 million in an all-cash deal .
0.67	< 0.01	the deal offers amicas_inc shareholders $ 5.35 for each share , a premium of 21 percent over the stock 's close on december 24 on nasdaq .
0.06	< 0.01	amicas_inc expects to close the transaction in the first quarter of 2010 , it said in a statement .
0.21	0.99	shares of the boston , massachusetts-based company closed at $ 4.42 thursday .
Our Method	**TD-HAN**	**Case**: No. 3; **Target**: Canon Inc.; CAR_3:4.7%; **Our method**:0.65; **TGT-CTX-LSTM**:0.71
0.05	0.01	**Abstract:** canon_inc (7751.t) posted a 30.9 percent decline in quarterly operating profit on monday , hurt by production halts due to parts shortages after the march 11 earthquake , but it raised its full-year forecast due to a faster recovery than expected .
0.2	0.03	canon_inc 's april-june figure came to 78.4 billion yen ($ 1 billion) , which is higher than an expected profit of 55.9 billion yen , the average of six analysts polled by thomson_reuters_corp i/b/e/s , but lower than the 113.4 billion yen it booked for the same quarter last year .
0.53	0.11	the world 's biggest maker of digital cameras also lifted its annual forecast to 380 billion yen , after slashed its full-year operating profit forecast following the devastating march earthquake and tsunami to 335 billion yen from 470 billion yen .
0.22	0.85	market expectations are for an annual profit of 365 billion yen , based on the average of 18 forecasts by analysts polled by thomson_reuters_corp i/b/e/s .

Table 3: Weights learned by our method and TD-HAN for each sentence in news documents. The actual CAR_3, the predicted probability of our method and TGT-CTX-LSTM are shown in the first row of each news document. The first sentence is the abstract.

6 Related Work

Our work falls into the area of mining financial text for stock market prediction. Rich linguistic features in financial text have been studied. Wang and Hua (2014) and Schumaker and Chen (2009) exploit shallow lexical and syntactic features, such as n-grams, noun-phrases and named entities. Such representations ignore the word orders and the important semantic relations between sentences, which thus can not fully represent the information conveyed in a document. With the rise of open information extraction techniques, Xie et al. (2013) exploit semantic frames and Ding et al. (2014) use event tuples to represent the news events. Such representation suffers from the problem of discreteness and data sparsity, which makes it difficult to generalize. The main difference between our method and this line of work is that they use discrete representations of the title or abstract of the financial news. Instead, we propose a fixed-length, continuous representations for the entire documents.

Our work also aligns with recent work on using deep neural models to learn embedding vectors for sentences and documents. Le and Mikilov (2014) extend the skip-gram algorithm (Mikolov et al., 2013) by regarding documents as contexts, training their embeddings together with training word embeddings. Kingma and Ba (2014) train sentence embeddings by using LSTMs to compose word vectors in sentences, leveraging neighbor sentences. Kenter et al. (2016) take a similar basis by using neighbor sentences to guide sentence vector training, but use a simpler network structure by calculating sentence embeddings as the sum of its word embeddings. Li et al. (2015) use a hierarchical auto-encoder structure to learn sentence and document embeddings. Yang et al. (2016) extends the hierarchical LSTM network of Li et al. (2015), applying attention for weighting different words and sentences, giving state-of-the-art accuracies for document classification. In contrast to the existing methods, we propose learning a target-specific representation of documents for a specific end task, namely news-driven market prediction.

7 Conclusion

We investigated target-specific document representations of financial news for cumulative abnormal return prediction in this paper, comparing a set of document embedding models. Empirical studies showed that a combination of target-specific news abstract representation and contextual sentence representation via the attention mechanism can give better results compared to several alternative sentence-level and document-level methods. Our final model demonstrated the usefulness of modeling a full document, as compared to only titles and abstracts, for obtaining more accurate results.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions to help improve this paper. This work was partly supported by the National Key Basic Research Program of China via grant 2014CB340503, the National Natural Science Foundation of China (NSFC) via grant 61472107 and 61702137.

References

Werner Antweiler and Murray Z Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *JCS*, 2(1):1–8.

Ching-Yun Chang, Yue Zhang, Zhiyang Teng Teng, Zahn Bozanic, and Bin Ke. 2016. Measuring the information content of financial news. In *26th Coling*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *EMNLP*, pages 1415–1425.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *IJCAI*, pages 2327–2333.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2133–2142.

Junwen Duan, Xiao Ding, and Ting Liu. 2018. Learning sentence representations over tree structures for target-dependent classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 551–560.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

SP Kothari and Jerold B Warner. 2004. The econometrics of event studies. *Available at SSRN 608601*.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1106–1115.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Robert P Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *TOIS*, 27(2):12.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1556–1566.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *ACL*, pages 1155–1165.

Boyi Xie, Rebecca J Passonneau, Leon Wu, and Germán G Creamer. 2013. Semantic frames to predict stock price movement. In *ACL*, pages 873–883.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.