Incorporating Multiple Knowledge Sources for Targeted Aspect-based Financial Sentiment Analysis

Anonymous ACL submission

Abstract

Combining symbolic and subsymbolic methods has become a promising strategy as research tasks in AI grow increasingly complicated and require a higher levels of understanding. Targeted Aspect-based Financial Sentiment Analysis (TABFSA) is one of such complicated tasks, as it involves information extraction, specification, and domain adaptation. External knowledge has been proven useful for general-purpose sentiment analysis, but not yet for the finance domain. Current state-of-the-art Financial Sentiment Analysis (FSA) models, however, have overlooked the importance of external knowledge. To fill this gap, we propose using attentive CNN and LSTM to strategically integrate multiple external knowledge sources into the pre-trained language model fine-tuning process for TABFSA. Experiments on the FiQA Task 1 and SemEval 2017 Task 5 datasets show that the knowledge-enabled models systematically improve upon their plain deep learning counterparts, and some outperform the ***state-of-the-art*** results reported in terms of aspect sentiment analysis error.

1 Introduction

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2015). The main objective of sentiment analysis is to classify the polarity of a given piece of text, which can be performed at the document-level (Moraes et al., 2013), sentence-level (Zhang and He, 2015), or aspect-level (Pontiki et al., 2016).

Early researches in FSA primarily focused on the document- or sentence-level polarity. However, it is *more common* for a single sentence to have multiple targets or aspects with different polarities for sentiment analysis of financial texts. Targeted Aspect-based Financial Sentiment Analysis (TABFSA), which aims to extract entities and aspects and detect their corresponding sentiment in financial texts, is thus a challenging but pragmatic task. The task involves target-aspect identification as well as polarity detection. Three examples of TABFSA are provided is Fig. 1.

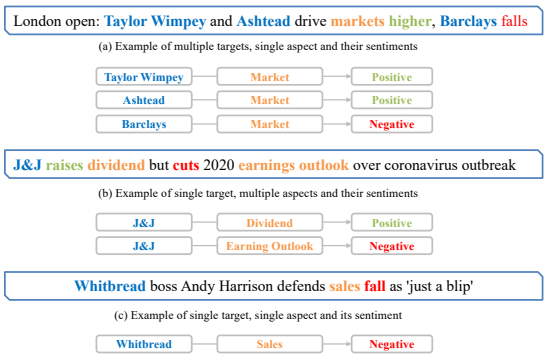

Figure 1: Example sentences with their target companies (blue), aspects (orange), and associated polarities detected.

For the two examples in (a) and (b), sentence-level sentiment analysis will assign a polarity value over the full text and mostly the opposite sentiment will nullify each other, resulting in a neutral overall sentiment. In contrast, TABFSA framework will provide positive sentiment to target "Taylor Wimpery" and "Ashtead" for the market aspect, and negative sentiment to target "Barclays" for the market aspect (Fig. 1). Similarly, a positive sentiment will be assigned to target "J&J" for the dividend aspect, but a negative sentiment for the earning outlook aspect.

There are two main sub-tasks for Targeted Aspect-based Sentiment Analysis (TABSA): the first sub-task is to extract aspects mentioned in the sentence, and the second one is to detect the sentiment for the corresponding targets and aspects. Generally, aspects can be extracted through frequency-based, syntax-based, unsupervised and supervised machine learning methods,

while sentiment polarity can be classified through lexicon-based or supervised machine learning approaches (Schouten and Frasincar, 2015). Recent methods may not require large-scale labelled data to generate predefined aspects. Instead, aspects are learned from a few keywords as supervision (Huang et al., 2020). The aspect extraction and sentiment detection sub-tasks could be performed either in a separate (Saeidi et al., 2016) or a joint manner (Wu and Ong, 2021).

TABSA has been studied and performed for a variety of domains such as movies, products, hotels, restaurants, healthcare, but it still remains much unexplored in the finance domain except for a few commercial products (Ho et al., 2019). We resort this observation to the following three reasons. *Firstly*, as elaborated by Araci (2019) and Liu et al. (2020), there is a lack of high-quality and large-scale open source finance domain-specific annotations. The research in fine-grained financial sentiment analysis has only gained more attention after the release of the "SemEval 2017 Task 5" and "FiQA Task 1" datasets. *Secondly*, lexical resources are limited and scattered. Since finance is a highly professional domain, general-purpose sentiment lexicons usually fail to take into account of the domain-specific connotations and the heavy reference to prior knowledge. For example, word like "liability" is considered negative in general-purpose sentiment analysis, but is frequent and has a neutral meaning in the financial context. This makes it difficult to generalize the sentiment classifiers and underlines the need for finance domain-specific sentiment analysis (Loughran and McDonald, 2011). *Lastly*, sentiment intensity scores are more consequential and nuanced for financial sentiment analysis compared to other domains. Whereas most of the current TABSA studies still adopt a polarity detection fashion (i.e. classification to positive or negative).

We propose knowledge-enabled (k-) transformer models to address the aforementioned challenges. In particular, our contributions can be summarized from three perspectives:

1. We adopted transformer models for the TABFSA task in a regressive fashion, and applied the most recent "recall-and-learn" strategy (Chen et al., 2020) to avoid the catastrophic forgetting problem.

2. We proposed using attentive CNN and LSTM to incorporate heterogeneous sentiment knowledge (both from domain-specific and general-purpose lexicons) into the fine-tuning process of pre-trained transformer models and demonstrated its effectiveness in complementing all the model training.

3. We achieved the best results to our knowledge over strong benchmark models on the two fine-grained financial sentiment analysis datasets, i.e., SemEval 2017 Task 5 and FiQA Task 1.

2 Related Work

2.1 Financial Sentiment Analysis

Sentiment analysis is proven to be a powerful tool for financial forecasting and decision making. The application scenarios include corporate disclosures, annual reports, earning calls, financial news, social media interactions and more (van de Kauter et al., 2015; Xing et al., 2018). Many interesting observations are reported, e.g., negative sentiment predicts short-term returns and volatility (Jiang et al., 2019; Xing et al., 2019), and strong sentiments for both directions seem to be more pronounced in fraudulent company reports (Hájek and Henriques, 2017).

Luo et al. (2018) categorized financial sentiment indicator into market-derived and human-annotated sentiments. The market-derived sentiments are computed from market dynamics, such as prices movement, trading volume, etc., thus may include noise from other sources. In this study, we investigate the subjective human-annotated sentiments, which are specifically labelled by professionals (Malo et al., 2014) or investors themselves (Xing et al., 2020). Instead of sentence-level sentiment polarity annotations, such as from the Financial PhraseBank (Malo et al., 2014), we focus on more fine-grained financial sentiment analysis datasets with targeted and targeted aspect-based sentiment intensity scores, i.e., SemEval 2017 Task 5 by Atzeni et al. (2017) and FiQA Task 1 by Maia et al. (2018) to the best of our knowledge. They are more useful for market prediction as opposite sentiment expressed in one news headlines for different targets tend to drive their market movement to opposite direction. In the remainder of this section, we review the many TABSA techniques experimented for these two datasets and their performance.

2.2 SemEval 2017 Task 5

This task has two separate tracks: both measure sentiment score predictions with cosine similarity.

In the *microblog messages* track, an ensemble of various boosted regressors based on linguistic features (Jiang et al., 2017) ranks first (Cosine=0.778), followed by another hybrid of deep learning and lexicon-based technique that combines Long Short-Term Memory (LSTM) and Convolution Neural Networks (CNN), proposed by Ghosal et al. (Cosine=0.751). In the *news headlines track*, CNN-based methods performed well. Later that same year, Akhtar et al. (2017) presented a method ensembling results generated from LSTM, CNN, GRU and SVM, using a Multi-Layer Perceptron (MLP), and achieved the state of the art for microblogs data (Cosine=0.797) and news headlines (Cosine=0.786).

2.3 FiQA Task 1

The FiQA (Financial Opinion mining and Question Answering challenge) Task 1, unlike SemEval 2017 Task 5, measures sentiment prediction performances mainly with mean squared error (MSE). de França Costa and da Silva (2018) established a strong baseline with a traditional feature engineering-based machine learning approach (MSE=0.0958). When target-aspect identification is jointly considered, biLSTM-CNN (Jangid et al., 2018) achieves an MSE of 0.112. This result is further pushed forward by an ensemble approach with an MSE of 0.0926 (Piao and Breslin, 2018).

In terms of pre-trained language models, ELMo (Embeddings from Language Models) and ULMFiT are suitable for TABFSA (Howard and Ruder, 2018; Peters et al., 2018). Yang et al. (2018) reported a good MSE of 0.08 using ULMFiT on the FiQA Task 1, whereas the best performance (MSE=0.07, R^2=0.55) is from a more recent fine-tuned language model: FinBERT (Araci, 2019).

3 Our Approach

Our study focuses on the sentiment detection part, not the target-aspect identification part of TABFSA, where the model is trained to predict the sentiment score given financial news headlines and posts, and their corresponding targets and aspects. Our aim being a comprehensive framework to utilize external knowledge, we search for an effective coupling of deep text representations and multiple knowledge sources individually developed.

3.1 Transformer Models

Although BERT is popular for TABSA (Araci, 2019; Wu and Ong, 2021), they would suffer from a pre-train fine-tuning discrepancy because the dependency between masked positions are neglected during the training phase. XLNet (Yang et al., 2019), which is an extension of Transformer-XL model on the other hand, can address this issue by using autoregressive method to learn the bidirectional contexts. Experiments show that XLNet has outperformed BERT in 20 NLP tasks significantly, including sentiment analysis and question answering. RoBERTa or Robustly optimized BERT approach, which is proposed by (Liu et al., 2019), is a replication study of BERT pre-training. It proposed an improved way of training BERT which includes (1) longer model training, with larger batches and byte-level Byte-Pair Encoding (BPE), over more data; (2) removal of the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training.

3.2 External Knowledge

External knowledge is demonstrated to be useful in many NLP tasks including TABSA. However, it is most commonly used as an input for RNNs (Ma et al., 2018; Bao et al., 2019) and CNNs (Shin et al., 2017), instead of used for the large pre-trained language model fine-tuning process. In most cases the knowledge is also single-sourced, hence does not deal with redundancy and contradiction problems.

Due to the lack of high-quality lexical resources for financial text, we incorporate multiple knowledge sources following three criteria: (1) both financial domain-specific and general-purpose lexicons are selected to balance precision and coverage; (2) the lexicons selected cover both sentiment and more fine-grained emotion knowledge; (3) lexicons that are created from social media text such as tweets and microblogs are purposely chosen for the sake of similar language style to our evaluation datasets. In sum, we consider 3 finance domain lexicons plus 6 general-purpose lexicons.

Finance domain lexicons consists of HFD, LM, and SMSL. **HFD** (Henry's Financial Dictionary) includes 104 positive words and 85 negative words, and is the first dictionary created specifically for the financial domain. It is used to measure the tone of earnings press releases, which are an important element of the firm-investor communication pro-

cess (Henry, 2008). HFD has been widely used for financial sentiment analysis. The weakness of HFD is its limited number of words and low coverage. **LM** (Loughran and McDonald) sentiment word list is created from the annual reports released by firms and includes 354 positive, 2,355 negative, 297 uncertainty, 904 litigious, 19 strong modal, 27 weak modal and 184 constraining words (Loughran and McDonald, 2011). LM is the most commonly used lexicon created for the finance domain that we are aware of. **SMSL** (Stock Market Sentiment Lexicon) is created from labeled StockTwits: tweets from a microblog platform specialized in stock market. SMSL includes 20,550 words and phrases and shows competitive results in measuring investor sentiments (Oliveira et al., 2016).

In addition, general-purpose lexicons are used to increase coverage, which comprise SenticNet (Cambria et al., 2020), VADER (Hutto and Gilbert, 2014), GI (Stone et al., 1966), NRC (Mohammad and Turney, 2013), OPL (Hu and Liu, 2004) and MPQA (Wilson et al., 2005).

3.3 Knowledge-enabled Transformer Models

Our model architecture is illustrated with Fig. 2. Specifically, BERT-base-cased (12-layer, 768-hidden, 12-heads, 109M parameters), XLNet-base-cased (12-layer, 768-hidden, 12-heads, 110M parameters) or RoBERTa-base (12-layer, 768-hidden, 12-heads, 125M parameters) is used to generate deep text representations. For the TABSA task, the input to BERT/XLNet/RoBERTa is a sentence pair. In consistent with (Sun et al., 2019), we denote a financial news headline or post as a sentence $S = \{x_1, x_2, ..., x_l\}$, and the auxiliary sentence containing the corresponding targets and aspects as $aux(S)$, e.g., "*what do you think of aspect for target?*" Then, the input is in a format of: "[cls] S [sep] $aux(S)$ [sep]" for BERT and RoBERTa and "S [sep] $aux(S)$ [sep][cls]" for XLNet. The output $\boldsymbol{U} \in \mathbb{R}^{768\times 1}$ is average pooled from the last hidden state.

In terms of external knowledge embedding, the nine selected lexicons are processed and formed as a *master dictionary*, where the key is a word or phrase, and the value is a list of associated sentiment and emotion scores. In our study, the master dictionary has 212,109 words and phrases, where each has 25 scores[1]. The scores are normalized to $[-1, +1]$, where -1 and $+1$ represent the most extreme sentiments. For each word $x_i \in S$, the external knowledge embedding $D(x_i)$ is looked up from such dictionary, and in case the word is not found, padded with zeros.

Because the original knowledge embedding uses lexicons individually developed, it contains inconsistent information and noise from alien language styles. To this end, we further apply feature selection techniques on training data to refine the most relevant knowledge for the learning process. We experimented with two popular methods to rank the feature importance, i.e., using random forest regressor (Farukee et al., 2020) or mutual information (Battiti, 1994; Sohangir et al., 2018). Random forest measures the mean decrease in impurity, while mutual information captures various kinds of dependency between variables, which is different from F-test that captures linear dependency only. We estimate mutual information based on entropy estimates from k-nearest neighbor distances ($k = 3$), as a larger k could introduce bias (Kraskov et al., 2004). Empirically, mutual information has produced better results: this aligns with the findings from (Frénay et al., 2013), where mutual information criterion is found to be able to choose features that minimize MSE and MAE in regressions.

The refined knowledge embedding for each sentence is $\boldsymbol{K}(S) \in \mathbb{R}^{m\times n}$, where $m > l$ is the maximum length of the sentences and n is the number of sentiment and emotion scores across lexicons[2]:

$$\boldsymbol{K}(S) = K_{x_1} \oplus K_{x_2} ... \oplus K_{x_l} \oplus \mathbf{0}^{(m-l)\times n}. \quad (1)$$

Inspired by Kim (2014) and Zhao and Wu (2016), for each word x_i we generate a context vector c_i using the attention layer to determine which word and lexicon should have more attention, and thus each sentence also has a contextual embedding $\boldsymbol{C}(S) \in \mathbb{R}^{m\times n}$.

$$c_i = \sum_{i\neq j} \alpha_{i,j} \cdot x_j \quad (2)$$

The attention weight $\alpha_{i,j}$ can be obtained by normalizing the score $s(x_i, x_j)$ from a MLP through softmax function, where given $k = |i - j| - 1$:

$$s(x_i, x_j) = (1 - \lambda)^k \cdot v_a^T tanh(W_a[x_i \oplus x_j]) \quad (3)$$

We concatenate the knowledge embedding with contextual embedding to form a 2-channel embed-

[1] Among those, 9 dimensions are contributed by SenticNet, 7 by NRC, and 1 by each else lexicon.

[2] In our case the optimal n ranges from $3 - 8$ out of the 25 dimensions.

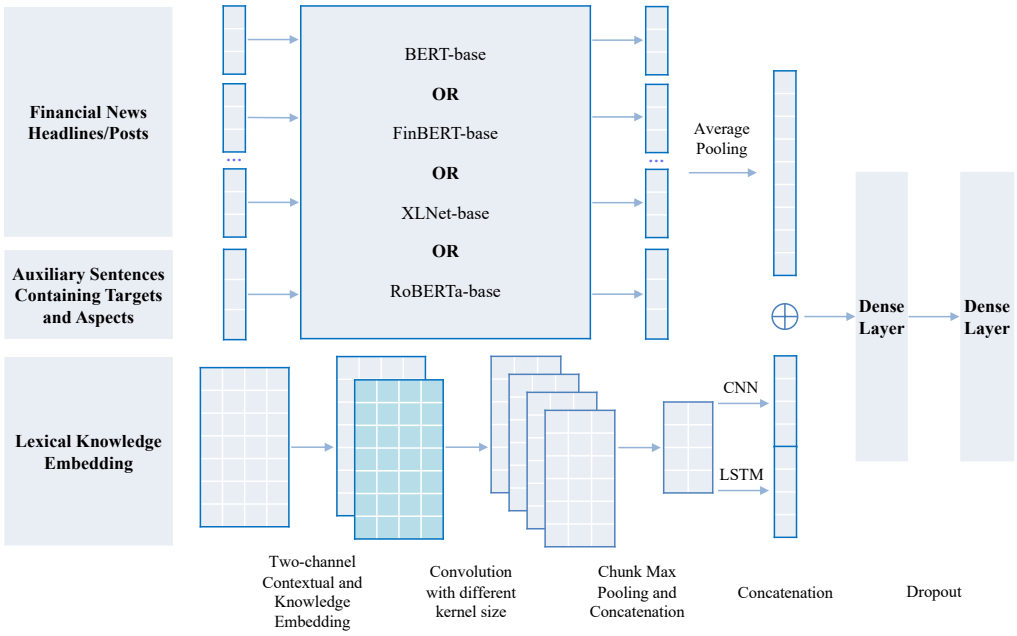

Figure 2: Architecture of the proposed knowledge-enabled transformer models.

ding and feed into CNNs to generate feature representation. The convolving kernel sizes are set to $d = (3,2), (3,3), (3,4), (3,5)$, and the number of filters c is experimented to be 4. Each convolution involves a filter $\boldsymbol{w} \in \mathbb{R}^{e \times h}$, where e is the number of lexicon scores and h is the number of words for a sliding window. A new feature z_{ij}, where $i \in [1, 2, ..., n-e+1]$ and $j \in [1, 2, ..., m-h+1]$, is generated from a sliding window over words $x_{(i:i+e-1, j:j+h-1)}$ as:

$$z_{ij} = w_{ij} \cdot K_{x_{(i:i+e-1, j:j+h-1)}} + b_{ij}, \quad (4)$$

where b_{ij} is a bias term.

The convolved feature vector $\boldsymbol{Z} \in \mathbb{R}^{(n-e+1)\times(m-h+1)}$ is represented by:

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & \cdots & z_{1(m-h+1)} \\ z_{21} & \cdots & z_{2(m-h+1)} \\ \vdots & \ddots & \vdots \\ z_{(n-e+1)1} & \cdots & z_{(n-e+1)(m-h+1)} \end{bmatrix}. \quad (5)$$

The convolved features $\boldsymbol{Z}$ are activated by ReLU function and global-max-pooled along i but chunk-max-pooled along j. The pooled feature maps are concatenated to form $\boldsymbol{P} \in \mathbb{R}^{c \times \sum p_i}$, where p_i is the length of pooled vector and $i \in [1, 2, 3, 4]$. Empirically, a second convolution is applied to $\boldsymbol{P}$ to further extract and downsize features, and in parallel LSTM is used to extract the sequential information. This way, we differ from Kim (2014) by using attentive convolution and chunk max pooling followed by additional CNN and LSTM layers in parallel to further extract features. The output $\boldsymbol{V} \in \mathbb{R}^{C_{out} \times 1}$ from CNN and $\boldsymbol{X} \in \mathbb{R}^{H_n \times 1}$ from LSTM are concatenated with $\boldsymbol{U}$ from a transformer model, where C_{out} is the number of channels produced by the last convolution and H_n is the dimension of the last hidden state of LSTM. Finally, this representation is passed through two linear layers with dropouts and sizes of $(768 + C_{out} + H_n, 768 + C_{out} + H_n)$, $(768 + C_{out} + H_n, 1)$. The output is therefore in a format of:

$$O = w_2 \cdot \sigma[w_1 \cdot \tanh(U \oplus V \oplus X) + b_1] + b_2, \quad (6)$$

where w_1, w_2, b_1, b_2 are weights and bias terms to be optimized with MSE loss.

3.4 The Catastrophic Forgetting Problem

One important issue in fine-tuning of pre-trained language models is the variability in error between different runs with the same configuration but different random seeds. Catastrophic forgetting and small training data size are two hypotheses for the origin of fine-tuning instability (Devlin et al., 2019; Dodge et al., 2020). To deal with the catastrophic

forgetting problem, Howard and Ruder (2018) proposed three training techniques: slanted triangular learning rates, discriminative fine-tuning, and gradual unfreezing. A more recent method (Chen et al., 2020), however, recalls knowledge from pre-training without the original data by using pre-training simulation mechanism and learns downstream tasks gradually by using objective shifting mechanism. This method achieved state-of-the-art results on our benchmark datasets. Therefore, we apply this "recall-and-learn" training strategy (Chen et al., 2020) to prevent catastrophic forgetting in our model fine-turning process for all pre-trained language models.

4 Experiments

4.1 Datasets

The SemEval 2017 Task 5 dataset is developed for fine-grained sentiment analysis on financial news and microblogs (Cortis et al., 2017). The training data includes 1,142 financial news headlines and 1,694 posts with their target entities and corresponding sentiment scores but without aspects labelled. The test data has 491 financial news headlines and 794 posts. The task is to extract and detect the targets and their corresponding sentiment scores. An example is shown in the textbox below.

```
"id": 2,
"company (target)": "Morrisons",
"title": "Morrisons book second consecutive
quarter of sales growth",
"sentiment": 0.43
```

The FiQA Task 1 dataset is from an open challenge (Maia et al., 2018), which consists of 498 financial news headlines and 675 posts with their target entities, aspects, and corresponding sentiment scores labeled. Although smaller than SemEval 2017 Task 5, FiQA Task 1 pre-defines four Level 1 aspects and 27 Level 2 aspects as shown in Table 1. The task, therefore, is to extract and detect both the targets, aspects, and their corresponding sentiment scores. Followed is one example from the FiQA Task 1 dataset.

Level 1	Level 2
Corporate	Reputation, Company Communication, Appointment, Financial, Regulatory, Sales, M&A, Legal, Dividend Policy, Risks, Rumors, Strategy
Stock	Options, IPO, Signal, Coverage, Fundamentals, Insider Activity, Price Action, Buyside, Technical Analysis
Economy	Trade, Central Banks
Market	Currency, Conditions, Market, Volatility

Table 1: Level 1 and Level 2 Aspects in FiQA dataset

```
"sentence": "Royal Mail chairman Donald Brydon
set to step down",
"info": [
"snippets": "['set to step down']",
"target": "Royal Mail",
"sentiment_score": "-0.374",
"aspects": "['Corporate/Appointment']" ]
```

4.2 Benchmarks

We benchmark our knowledge-enabled models with plain BERT-base-cased, FinBERT-base-cased, XLNet-base-cased and RoBERTa-base models. BERT variants (Sun et al., 2019; Xu et al., 2019; Wu and Ong, 2021) are chosen because many are developed for the (T)ABSA task and some achieved state-of-the-art results. Moreover, FinBERT (Araci, 2019) performs further pre-training to address the domain-specific language style and was ranked the first for sentiment analysis on Financial PhraseBank[3]. Similarly, the input to the pre-trained language model is a sentence pair, in which one sentence is the auxiliary sentence containing the target and aspect and the other sentence is the financial news headline or post. The average pooling is performed on the last hidden state followed by dropout. A last linear layer is added with the size of (768×1). Loss function minimizes MSE.

4.3 Other Experimental Details

The FiQA Task 1 dataset is split into 90% for training and 10% for test by performing 10-fold split. The validation dataset, which is 25% of the training data, is used to select the best model and the test dataset is used to report the final performance scores. Since gold standard is not released, we perform 10-fold cross-validation on two differently-seeded runs for evaluation, and the mean score is reported. As for SemEval 2017 Task 5 dataset, it is split into 75% training and 25% validation to train the model 10 times with different random seeds and the gold standard dataset is used to report the mean performance score. Our models are configured and trained on an NVIDIA Tesla-P100-PCIe-16GB processor with 100 epochs, learning rate of 3e-5, and Recall Adam as the optimizer.

[3]https://paperswithcode.com/sota/sentiment-analysis-on-financial-phrasebank

Model	Headline		Post	
	Cosine	R^2	Cosine	R^2
(Jiang et al., 2017)	0.7100	-	0.7780	-
(Akhtar et al., 2017)	0.7860	-	0.7970	-
k-BERT	0.7969	0.635	0.7912	0.586
k-FinBERT	0.8009	0.642	0.7853	0.576
k-XLNet	0.8270	0.685	0.8074	0.615
k-RoBERTa	**0.8483**	**0.721**	**0.8126**	**0.624**

Table 2: Performance of proposed knowledge-enabled transformer models in comparison to the state-of-the-art approaches on SemEval 2017 Task 5. Boldface indicated the best result. We use the results reported in Jiang et al. (2017) and Akhtar et al. (2017). "-" means not reported.

Model	MSE	R^2
(Piao and Breslin, 2018)	0.0926	0.414
(Yang et al., 2018)	0.0800	0.400
(Araci, 2019)	0.0700	0.550
k-BERT	0.0628	0.615
k-FinBERT	0.0646	0.603
k-XLNet	0.0532	0.674
k-RoBERTa	**0.0490**	**0.711**

Table 3: Performance of proposed knowledge-enabled transformer models in comparison to the state-of-the-art approaches on FiQA Task 1. Boldface indicated the best result. We use the results reported in Piao and Breslin (2018), Yang et al. (2018) and Araci (2019)

5 Results & Analysis

In consistent with previous studies, we respectively report cosine similarities for SemEval 2017 Task 5, and MSE for FiQA Task 1 (see Table 2 and Table 3). Additionally, R^2 measures the percentage of variance explained by the model under evaluation. Under all columns and metrics, FinBERT outperforms BERT in news headline data, attesting to the effectiveness of domain-specific pre-training. Notably, XLNet and RoBERTa outperform BERT and FinBERT by significant margins even before the integration of sentiment knowledge. This confirms RoBERTa and XLNet as a more effective deep representation model than BERT for the TABFSA task. Knowledge-enabled RoBERTa achieves state-of-the-art results on both SemEval 2017 Task 5 (Cosine[h]=0.8483, Cosine[p]=0.8126) and FiQA Task 1 (MSE= 0.0490, R^2=0.711). Those metrics are circa 5% improvement from the previous best results on SemEval 2017 Task 5 by Akhtar et al. (Cosine[h]=0.7860, Cosine[p]=0.7970), and circa 30% improvement from the previous best results on FiQA Task 1 by Araci (MSE= 0.07, R^2=0.55).

Cosine Similarity	Headline			Post		
	Mean	Median	SD	Mean	Median	SD
BERT	0.7935	0.7904	0.0096	0.7886	0.7850	0.0108
k-BERT	**0.7969**	**0.7958**	**0.0072**	**0.7912**	**0.7932**	**0.0104**
FinBERT	0.7969	0.7987	0.0093	0.7817	0.7823	**0.0093**
k-FinBERT	**0.8009**	**0.8019**	**0.0069**	**0.7853**	**0.7839**	0.0105
XLNet	0.8199	0.8186	0.0151	0.8031	0.8025	0.0110
k-XLNet	**0.8270**	**0.8261**	**0.0091**	**0.8074**	**0.8098**	**0.0090**
RoBERTa	0.8430	0.8423	**0.0080**	0.8085	0.8082	0.0136
k-RoBERTa	**0.8483**	**0.8500**	0.0170	**0.8126**	**0.8118**	**0.0125**

Table 4: Ablation Analysis for SemEval 2017 Task 5.

MSE	Mean	Median	SD
BERT	0.0651	0.0602	0.0191
k-BERT	**0.0628**	**0.0573**	**0.0180**
FinBERT	0.0675	0.0668	0.0172
k-FinBERT	**0.0646**	**0.0623**	**0.0157**
XLNet	0.0549	0.0526	0.0147
k-XLNet	**0.0532**	**0.0502**	**0.0119**
RoBERTa	0.0548	0.0526	**0.0173**
k-RoBERTa	**0.0490**	**0.0420**	0.0185

Table 5: Ablation Analysis for FiQA Task 1.

5.1 Ablation Analysis

Ablation analysis is performed to validate the external knowledge embedding module. The results of models trained with different random seeds for various transformer models and knowledge-enabled transformer models are provided in Table 4 and Table 5, which shows the positive impact of knowledge integration on model performance and stability.

It is observed that the integration of external knowledge has improved both accuracy and stability across benchmark models. Specifically, the FiQA Task 1 data has reported a 4% improvement in MSE for BERT and FinBERT with smaller standard deviation (SD). The knowledge-enabled RoBERTa has decreased the MSE by 10% from 0.0548 to 0.0490 although the model is destabilized slightly. As for SemEval 2017 Task 5, the knowledge-enabled RoBERTa has improved cosine similarities from 0.8430 to 0.8483 for headline data and from 0.8085 to 0.8126 for post data.

5.2 Knowledge Quality Analysis

The precision and coverage of lexicons impact the effectiveness of external knowledge integration into the fine-tuning process. It is observed that with the increase in the number of lexicon scores incorporated by mutual information, the model per-

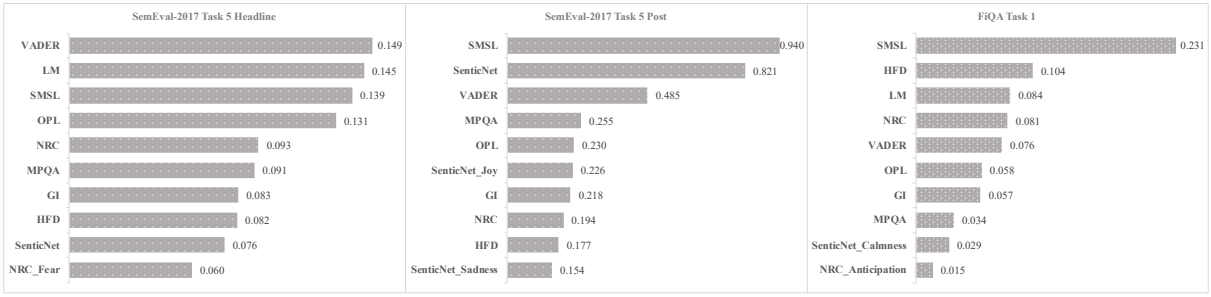

Figure 3: Mutual information of lexicons for SemEval 2017 Task 5 and FiQA Task 1 datasets.

formance initially increases but subsequently fluctuates or even decreases, which means relevant knowledge is able to improve the model performance but noise knowledge will potentially destabilize the model. There is a balance in sufficiency and redundancy of knowledge to ensure the right coverage and precision to compliment learning process. Moreover, the closer to the accuracy bound of the deep neural network, the more difficult to improve the results by including external knowledge.

We discover that the optimal dimension of lexicon scores ranges from 3 to 8, and their mutual information can be used to rank and pre-select relevant knowledge (see Fig. 3). In terms of lexicon scores selected, the experiment shows that both sentiment and emotion knowledge are helpful, though generally sentiment scores are more important than emotion scores. The importance of finance domain-specific lexicons such as LM (Loughran and McDonald, 2011) and SMSL (Oliveira et al., 2016) are consistently higher than general-purpose lexicons. As for emotion dimensions, joy, sadness and fear tend to be more relevant for the TABFSA task.

5.3 Error Analysis

Although in the majority of cases, incorporating external knowledge is beneficial for the accuracies of predicted sentiment scores, we also observed errors in some cases. We describe these two scenarios by comparing sentiment scores predicted by RoBERTa and knowledge-enabled RoBERTa.

Scenario 1

Sentence: $NKE gapping up to all time highs

Sentiment_Ground_Truth: 0.782

Sentiment_RoBERTa: 0.468

Sentiment_knowledge-enabled RoBERTa: 0.603

Lexicon_score_sum: [0.3, 0, 2.0]

In Scenario 1, knowledge-enabled RoBERTa has improved the sentiment score significantly from 0.468 to 0.603, as words such as ‘up’ and ‘high’ are consistently positive in the selected lexicons, which results in a strongly positive tone as shown in the sum of selected lexicon scores from SMSL, LM and HFD.

In Scenario 2, however, knowledge-enabled RoBERTa is no better than the standalone RoBERTa. For this concrete example, although the sippet of “invalidated by US court” is negative, the word ‘invalidated’ does not carry any sentiment in 2 out of the 3 selected lexicons. While ‘patent’, ‘drug’ and ‘court’ are positive words in SMSL, leading the overall sentiment prediction to a more neutral score. The sentiment of words ‘drug’ and ‘court’ in this context is considered aforementioned noise knowledge.

Scenario 2

Sentence: AstraZeneca’s patent on asthma drug invalidated by US court

Sentiment_Ground_Truth: -0.656

Sentiment_RoBERTa: -0.392

Sentiment_knowledge-enabled RoBERTa: -0.252

Lexicon_score_sum: [0.87, -1, 0.0]

6 Conclusion and Future Work

A framework that strategically combines symbolic (heterogeneous sentiment lexicons) and subsymbolic (deep language model) modules for TABFSA is proposed in this research. Specifically, we are pioneering in employing attentive CNN and LSTM to touch multiple knowledge sources and integrating with transformer models. The incorporation of external knowledge into transformer models has achieved state-of-the-art performance on both the SemEval 2017 Task 5 and the FiQA Task 1 datasets. We plan to investigate three further issues in future work: 1) influence of domain-specific lexicon coverage on their effectiveness, 2) alternative methods for knowledge embedding, and 3) what affects the effectiveness of different transformer architecture, e.g., RoBERTa vs. XLNet.

References

Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *EMNLP*, pages 540–546.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Mattia Atzeni, Amna Dridi, and Diego Reforgiato Recupero. 2017. Fine-grained sentiment analysis on financial microblogs and news headlines. In *Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017*, pages 124–128.

Lingxian Bao, Patrik Lambert, and Toni Badia. 2019. Attention and lexicon regularized lstm for aspect-based sentiment analysis. In *Proceedings of ACL: Student Research Workshop*, pages 253–259.

Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550.

Erik Cambria, Yang Li, Frank Xing, Soujanya Poria, and Kenneth Kwok. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 105–114.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *EMNLP*, pages 7870–7881.

Keith Cortis, Andre Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *International Workshop on Semantic Evaluation*.

Dayan de França Costa and Nadia Felix Felipe da Silva. 2018. Inf-ufg at fiqa 2018 task 1: predicting sentiments and aspects on financial tweets and news headlines. In *Companion Proceedings of the The Web Conference 2018*, pages 1967–1971.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Minhaz Bin Farukee, MS Zaman Shabit, Md Rakibul Haque, and AHM Sarowar Sattar. 2020. Ddos attack detection in iot networks using deep learning models combined with random forest as feature selector. In *International Conference on Advances in Cyber Security*, pages 118–134.

Benoît Frénay, Gauthier Doquire, and Michel Verleysen. 2013. Is mutual information adequate for feature selection in regression? *Neural Networks*, 48:1–7.

Deepanway Ghosal, Shobhit Bhatnagar, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Iitp at semeval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In *International Workshop on Semantic Evaluation*, pages 899–903.

Petr Hájek and Roberto Henriques. 2017. Mining corporate annual reports for intelligent detection of financial statement fraud - A comparative study of machine learning methods. *Knowledge Based Systems*, 128:139–152.

Elaine Henry. 2008. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4):363–407.

Shuk Ying Ho, Ka Wai (Stanley) Choi, and Fan (Finn) Yang. 2019. Harnessing aspect-based sentiment analysis: How are tweets associated with forecast accuracy? *Journal of the Association for Information Systems*, 20(8):1174–1209.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*, pages 328–339.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *EMNLP*, pages 6989–6999.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Aspect-based financial sentiment analysis using deep learning. In *Companion Proceedings of the The Web Conference 2018*, pages 1961–1966.

Fuwei Jiang, Joshua Lee, Xiumin Martin, and Guofu Zhou. 2019. Manager sentiment and stock returns. *Journal of Financial Economics*, 132:126–149.

Mengxiao Jiang, Man Lan, and Yuanbin Wu. 2017. Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In

International Workshop on Semantic Evaluation, pages 888–893.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E*, 69(6):066138.

Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Yinhan Liu, 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. *arXiv preprint arXiv:1907.11692*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4513–4519.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, pages 4244–4250.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *AAAI*, pages 5876–5883.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Rodrigo Moraes, João Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.

Guangyuan Piao and John G Breslin. 2018. Financial aspect and sentiment predictions with deep neural networks: an ensemble approach. In *Companion Proceedings of the The Web Conference 2018*, pages 1973–1977.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, and al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*, pages 19–30.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *COLING*, pages 1546–1556.

Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 28(3):813–830.

Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, Denmark. Association for Computational Linguistics.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT*, pages 380–385.

Marjan van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11):4999–5010.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, pages 347–354.

Zhengxuan Wu and Desmond C. Ong. 2021. Context-guided BERT for targeted aspect-based sentiment analysis. In *AAAI*, pages 14094–14102.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *COLING*, pages 978–987.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Frank Z. Xing, Erik Cambria, and Yue Zhang. 2019. Sentiment-aware volatility forecasting. *Knowledge Based Systems*, 176:68–76.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of NAACL-HLT*, pages 2324–2335.

Steve Yang, Jason Rosenfeld, and Jacques Makutonin. 2018. Financial aspect-based sentiment analysis using deep representations. *arXiv preprint arXiv:1808.07931*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.

Pu Zhang and Zhongshi He. 2015. Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science*, 41(4):531–549.

Zhiwei Zhao and Youzheng Wu. 2016. Attention-based convolutional neural networks for sentence classification. In *INTERSPEECH*, pages 705–709.