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

KELVIN DU, School of Computer Science and Engineering, Nanyang Technological University, Singapore
FRANK XING, Department of Information Systems and Analytics, National University of Singapore, Singapore
ERIK CAMBRIA, School of Computer Science and Engineering, Nanyang Technological University, Singapore

Combining symbolic and subsymbolic methods has become a promising strategy as research tasks in AI grow increasingly complicated and require higher levels of understanding. Targeted Aspect-based Financial Sentiment Analysis (TABFSA) is an example of such complicated tasks, as it involves processes like information extraction, information specification, and domain adaptation. However, little is known about the design principles of such hybrid models leveraging external lexical knowledge. To fill this gap, we **define** anterior, parallel, and posterior knowledge integration and **propose** incorporating multiple lexical knowledge sources strategically into the fine-tuning process of pre-trained transformer models for TABFSA. Experiments on the Financial Opinion mining and Question Answering challenge (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 **state-of-the-art** results reported in terms of aspect sentiment analysis error. We discover that parallel knowledge integration is the most effective and domain-specific lexical knowledge is more important according to our ablation analysis.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Neural networks**; • **Information systems** → *Information retrieval;*

Additional Key Words and Phrases: Financial sentiment analysis, neural networks, knowledge enabled system, deep learning, transformer models

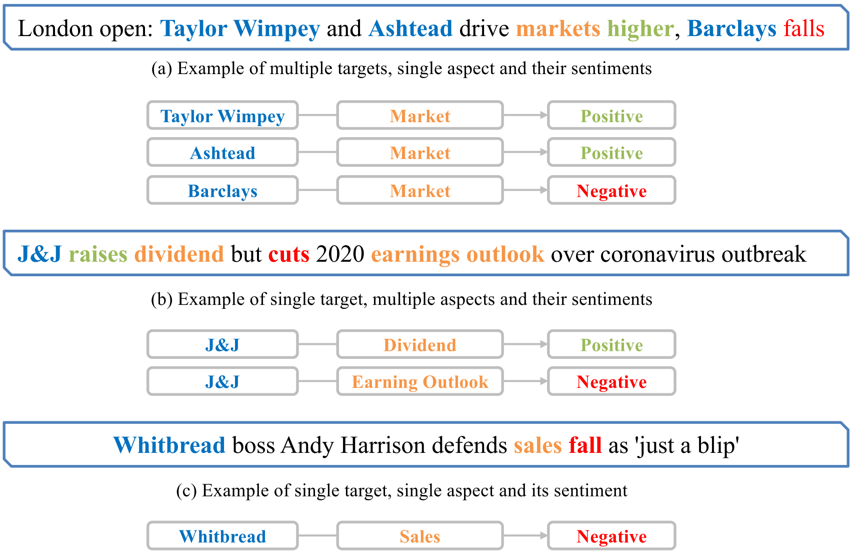
ACM Reference format:
Kelvin Du, Frank Xing, and Erik Cambria. 2023. Incorporating Multiple Knowledge Sources for Targeted Aspect-based Financial Sentiment Analysis. *ACM Trans. Manag. Inform. Syst.* 14, 3, Article 23 (June 2023), 24 pages.
https://doi.org/10.1145/3580480

1 INTRODUCTION

Sentiment analysis analyzes people's sentiments, attitudes, opinions, emotions, evaluations, and appraisals toward various entities such as events, topics, services, products, individuals, organizations, issues, and their attributes [38]. The main objective of sentiment analysis is to classify the

Authors' addresses: K. Du and E. Cambria, School of Computer Science and Engineering, Nanyang Technological University, Block N4 #02a-32, Nanyang Avenue, Singapore 639798; emails: zidong001@e.ntu.edu.sg, cambria@ntu.edu.sg; F. Xing, Department of Information Systems and Analytics, National University of Singapore, 13 Computing Drive, Singapore 117417; email: xing@nus.edu.sg.

2158-656X/2023/06-ART23 $15.00
https://doi.org/10.1145/3580480


London open: Taylor Wimpey and Ashtead drive markets higher, Barclays falls
Taylor Wimpey
Market
Positive
Ashtead
Market
Positive
Barclays
Market
Negative
J&J raises dividend but cuts 2020 earnings outlook over coronavirus outbreak
J&J
Dividend
Positive
J&J
Earning Outlook
Negative
Whitbread boss Andy Harrison defends sales fall as 'just a blip'
Whitbread
Sales
Negative

(a) Example of multiple targets, single aspect and their sentiments

(b) Example of single target, multiple aspects and their sentiments

(c) Example of single target, single aspect and its sentiment

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

polarity of a given piece of text, which can be performed at the document [16, 56], sentence [53, 81], or aspect level [24, 50]. The objective can be more challenging when sentiment analysis is applied to professional language domains, such as finance [45]. Early research in **Financial Sentiment Analysis** (**FSA**) primarily focused on the document- or sentence-level sentiment polarities. 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 in Figure 1.

For the two examples in Figure 1(a) and (b), sentence-level sentiment analysis will assign a polarity value over the complete text, and mostly the opposite sentiment will nullify each other, resulting in overall neutral sentiment. In contrast, the 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 (Figure 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. From this perspective, TABFSA has its practical significance. There is a great deal of professionalism involved in finance and it is vital that the information used in finance is accurate and precise. Otherwise, the wrong decision would be made that could result in economic losses. TABFSA can enhance the quality of financial sentiment analysis, which is critical for downstream applications, such as financial forecasting and financial decision-making. It is common for two entities to have opposite sentiments in one sentence, for example. In this case, a market prediction based on sentence-level sentiment is inaccurate, but TABFSA can address this issue and extract sentiment for each entity for subsequent market predictions.

There are two main sub-tasks for TABSA: the first sub-task is to extract aspects mentioned in the sentence, and the second 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 [62]. Current methods may not require large-scale labeled data to generate predefined aspects. Instead, aspects are learned from a few keywords as supervision [29]. The aspect extraction and sentiment detection sub-tasks could be performed either in a separate [61] or in a joint manner [70].

TABSA has been studied and performed for various domains such as movies, products, hotels, restaurants, and healthcare, but it remains much unexplored in the finance domain except for a few commercial products [26]. We attribute this observation to the following three reasons. *Firstly*, as previous literature has pointed out [2, 41], 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" (**Financial Opinion mining and Question Answering challenge**) datasets. *Secondly*, lexical resources are limited and scattered. Since finance is a highly professional domain, general-purpose sentiment lexicons usually fail to consider the domain-specific connotations and the heavy reference to prior knowledge. For example, a 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 [43]. *Lastly*, sentiment intensity scores are more consequential and nuanced for financial sentiment analysis than 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, which aims to answer the following research questions:

(1) Can integration of lexical knowledge improve the performance of pre-trained language models in TABFSA tasks?
(2) The methods to integrate knowledge into the fine-tuning process of pre-trained language models can be generally categorized into three types: anterior, parallel, and posterior integration. When multiple sources of lexical knowledge are provided, among anterior, parallel, or posterior integration, which is a more effective approach to incorporate knowledge?
(3) To improve the domain application of pre-trained language models, one method adopted by researchers is to train domain-specific pre-trained language models such as FinBERT, but it requires a large domain-specific corpus and considerable computing resources. Does incorporation of financial knowledge produce better model performance than retraining of finance domain-specific language models in TABFSA task?

In particular, our contributions can be summarized from four perspectives:

(1) We defined ***anterior, parallel, and posterior*** knowledge integration and conducted extensive experiments to examine the best approach to incorporate multiple lexicon knowledge into the fine-tuning process of transformer models and identified that the parallel approach is more effective in combining multiple lexical knowledge sources and pre-trained language models.
(2) We proposed incorporating heterogeneous sentiment knowledge (both from domain-specific and from 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 demonstrated that the incorporation of lexical knowledge produces better model performance than retraining of finance domain-specific language models in TABFSA. The lack of knowledge in the FSA task makes knowledge integration valuable. 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

FSA is 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 [68, 71, 74]. Many exciting observations have been reported, e.g., negative sentiment predicts short-term returns and volatility [32, 72], and strong sentiments for both directions seem to be more pronounced in fraudulent company reports [22].

Luo et al. [44] categorized financial sentiment indicator into market-derived and human-annotated sentiments. The market-derived sentiments were computed from market dynamics, such as price movement and trading volume, thus may include noise from other sources. In this study, we investigate the subjective human-annotated sentiments, which were specifically labeled by professionals [48] or investors themselves [73]. Instead of sentence-level sentiment polarity annotations, such as from the Financial PhraseBank [48], 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 [3] and FiQA Task 1 by [47], to the best of our knowledge. They are more useful for market predictions as the opposite sentiment expressed in one news headline for different targets tends to drive their market movement to the opposite direction. In the remainder of this section, we review TABSA techniques and their performances, experimented with these two datasets, including lexicon-based [75], machine learning–based or deep learning–based methods, and hybrid methods.

2.2 SemEval 2017 Task 5

The SemEval 2017 Task 5 contains two tracks: News Statements/Headlines and Microblog Messages. The news headlines were crawled from difference sources such as Yahoo Finance[1] and the microblog messages were obtained from StockTwits[2] and Twitter.[3] The evaluation of sentiment score prediction is based on weighted cosine similarity, which aims to compare the proximity between predicted results and gold standard. The FSA techniques in earlier studies include lexicon-based, machine learning–based, deep learning–based methods, and hybrid methods. Lexicon-based methods detect the sentiment of the text by analyzing the semantic orientation of the words in the text. For example, the general-purpose lexicon-based sentiment analyzer includes TextBlob [42], SnowNLP,[4] and SentiWordNet [4]. Machine learning approaches construct features and use classification or regression algorithms to determine sentiment. In contrast, deep learning approaches construct complicated representations from textual data with a high level of abstraction, using the **Convolutional Neural Networks** (**CNNs**) and **Recurrent Neural Networks** (**RNNs**) and their variants and have achieved remarkable performances in sentiment analysis. In the *microblog messages* track, an ensemble of various regressors (i.e., AdaBoost, Bagging, Random Forest, Gradient Boosting, LASSO, Support Vector Regression, and XGBoost) based on linguistic, sentiment lexicon, domain-specific features, and word embeddings [33] ranks first (Cosine = 0.778), followed by a hybrid of deep learning and lexicon-based technique that combined CNN, LSTM, feature-driven **Multi-Layer Perceptron** (**MLP**), and vector-averaging MLP, proposed by [20] (Cosine = 0.751). In the *news headlines track*, CNN-based methods performed well. The highest score (Cosine = 0.745) was reported by [49], which combined GloVe [58] and DepecheMood [65] to represent words and fed into CNN

[1] http://finance.yahoo.com/.
[2] http://stocktwits.com/.
[3] https://twitter.com.
[4] https://pypi.org/project/snownlp/.

followed by global max-pooling. The output was then concatenated with VADER [30] sentiment scores for two levels of dropout and fully connected layers. [34] combined the representations learned from CNN and bidirectional **Gated Recurrent Unit** (**GRU**) with an attention mechanism with hand-engineered lexical, sentimental, and metadata features, and obtained weighted cosine similarity scores of 0.723 and 0.744 for Microblogs and News Headline tracks, respectively. Later, [1] presented a method ensembling results generated from LSTM, CNN, GRU, and SVM, using a MLP, and achieved the state-of-the-art performance for microblogs data (Cosine = 0.797) and news headlines (Cosine = 0.786). Recently, MetaPro [51] was proposed to improve financial news headline sentiment analysis by identifying and interpreting metaphors. Metaphors commonly appear in the financial text, causing errors for classifiers. By paraphrasing metaphors into their literal counterparts, three state-of-the-art sentiment classifiers achieved large improvements.

2.3 FiQA Task 1

The FiQA Task 1 measures sentiment prediction performances mainly with **Mean Squared Error** (**MSE**) and aspect identification with F1-Score, which is different from SemEval 2017 Task 5. Since FiQA Task 1 provided both target and aspect labels, in addition to machine learning and deep learning methods, many hybrid-based and pre-trained language models were also proposed to solve the target-aspect identification problem. In particular, [12] established a strong baseline with a traditional feature engineering-based machine learning approach (MSE = 0.0958) by treating aspect extraction as a classification task and sentiment detection as a regression task, using **Support Vector Classifier** (**SVC**) and **Support Vector Regressor** (**SVR**), respectively. The generated features included n-gram, tokenization, word replacements, and word embeddings using Word2Vec and **Term Frequency–Inverse Document Frequency** (**TF-IDF**). When target-aspect identification is jointly considered, the biLSTM-CNN proposed by [31] treated aspect extraction as a multiclass classification problem, as this task does not involve multiple aspects for one target. First, it adopted bidirectional LSTM to extract aspects using word embeddings such as GloVe [58], Google-News-Word2Vec [54], Godin [21], FastText [7], and Keras[5] in-built embedding layer. Meanwhile, a multi-channel CNN is used for sentiment analysis task with enhanced vector combined from the dependency tree, sentence word vector, and snippet and target vector. The Bayesian optimization was used for hyperparameters tuning to find out the most optimal parameters. This method achieved 0.69 F1 for aspect extraction and 0.112 MSE for sentiment analysis. This result was further pushed forward by an ensemble approach with an MSE of 0.0926 [60]. This method ensembled CNNs and RNNs with a voting strategy and a ridge regression for aspect and sentiment predictions.

Regarding pre-trained language models, **embeddings from Language Models** (**ELMo**) [59] and ULMFiT [27] are suitable for TABFSA. Yang et al. [78] reported a good MSE of 0.08, using ULMFiT on the FiQA Task 1. With ULMFiT, users can transfer word representations with vectors and use a single pre-trained model architecture called AWD-LSTM for all intended tasks. ULMFiT differs from ELMo, in that ELMo requires users to concatenate the outputs of each trained layer ultimately and then use the resulting fixed embeddings to perform downstream tasks, while ULMFiT fine-tunes a whole language model to its target domain, then connects it directly to downstream tasks. A more recent fine-tuned language model FinBERT [2] reported the best performance (MSE = 0.07, R^2 = 0.55) on the FiQA Task 1.

2.4 Knowledge Incorporation

The incorporation of lexical knowledge is demonstrated to be useful in sentiment analysis [9, 37] and low-resource learning tasks [52]. However, it was most commonly used as an input for

[5] https://keras.io/layers/embeddings/.

RNNs [5, 46] and CNNs [63] but not much explored for the large pre-trained language model fine-tuning process. Although BERT-like pre-trained language models are capable of capturing general language representations, they lack domain-specific knowledge [39]. To improve the domain application of pre-trained language models, researchers have attempted to pre-train domain-specific language models such as FinBERT [2, 41], although the process requires a large domain-specific corpus and significant computing resources. Concurrently, other research has been conducted to study the techniques to incorporate domain-specific knowledge, such as knowledge graphs, into the pre-trained language model fine-tuning process in recent years [23]. For example, [39] integrated knowledge graph, a hybrid language and knowledge resource, namely, HowNet [15], and a never-ending Chinese knowledge extraction system called CN-DBPedia [76], by introducing a sentence tree for sentiment analysis, semantic similarity, **Question & Answering** (**Q&A**), and **Named Entity Recognition** (**NER**) tasks. The results show that the selected knowledge graphs have no significant effect on the sentiment analysis but better performance in semantic similarity, Q&A, and NER tasks. To improve the performance of aspect-based sentiment analysis, [82] leveraged sentiment knowledge graph to provide external domain knowledge for BERT. Nevertheless, few of the earlier researches attempted to incorporate lexicon knowledge, e.g., a common knowledge base available for many domains, into pre-trained language models. In most cases, the knowledge is also single-sourced and hence does not deal with redundancy and contradiction problems.

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. With the 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 famous for TABSA [2, 70], they would suffer from a pre-train fine-tuning discrepancy because the dependency between masked positions is neglected during the training phase. XLNet [80], which is an extension of the Transformer-XL model, on the other hand, can address this issue by using an autoregressive method to learn the bidirectional contexts. Experiments show that XLNet has significantly outperformed BERT in 20 NLP tasks, including sentiment analysis and question answering. RoBERTa or Robustly optimized BERT approach is another widely used pre-trained language model [19], which is proposed by [40], and is a replication study of BERT pre-training. It proposed an improved way of training BERT that includes (1) longer model training, with larger batches and byte-level **Byte-Pair Encoding** (**BPE**), over more data; (2) training on longer sequences; (3) removal of the next sentence prediction objective; and (4) changing the masking pattern applied to the training dynamically. In our study, we have adopted BERT, XLNet, and RoBERTa as baseline models to examine the effectiveness of knowledge incorporation into transformer models.

3.2 Lexical Knowledge

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; and (3) lexicons that are created from social media text such as tweets and microblogs are purposely chosen for the sake of similar

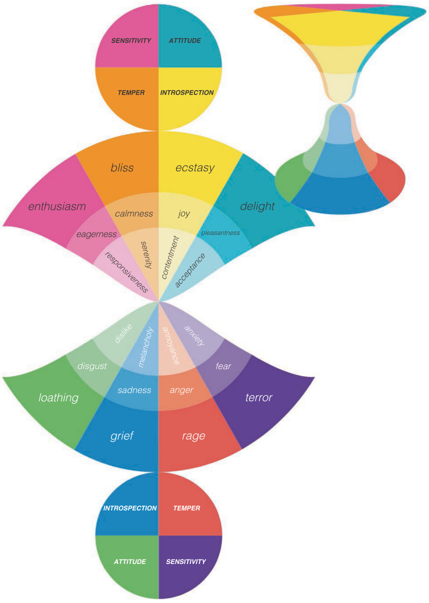


Fig. 2. Emotion dimensions included in SenticNet.

language style to our evaluation datasets. In sum, we consider three finance domain lexicons plus six general-purpose lexicons.

Finance domain lexicons consist of HFD, LM, and SMSL. **HFD** (**Henry's Financial Dictionary**) includes 104 positive and 85 negative words, and is the first dictionary explicitly created for the financial domain. It is used to measure the tone of earnings press releases, which are an essential element of the firm-investor communication process [25]. HFD has been widely used for financial sentiment analysis. The weakness of HFD is its limited number of words and low coverage. The **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 [43]. To our knowledge, LM is the most commonly used lexicon created for the finance domain. **SMSL** (**Stock Market Sentiment Lexicon**) is created from labeled StockTwits: tweets from a microblogging platform specialized in the stock market. SMSL includes 20,550 words and phrases and shows competitive results in measuring investor sentiments [57].

In addition, general-purpose lexicons are used to increase coverage, which comprise SenticNet [8], VADER [30], GI [66], NRC [55], OPL [28], and MPQA [69]. **SenticNet** is a general-purpose sentiment knowledge base with 200,000 commonsense concepts in its latest version [8]. Each concept in SenticNet is associated with rich dimensional emotion information (see Figure 2). For example, in record `senticnet[’sales_fall’] = [’-0.78’, ’-0.84’, ’0’, ’0’, ’#sadness’, ’#anger’, ’negative’, ’-0.81’]`, “−0.78” is the introspection (joy-versus-sadness) value, “” is the temper (calmness-versus-anger) value, “0”s are the attitude (pleasantness-versus-disgust) and sensitivity (eagerness-versus-fear) values, #sadness” is the primary mood, “#anger” is the secondary mood, “negative” is the polarity label, and “−0.81” is the polarity value. **VADER** is specifically tailored for sentiments expressed in social media [30]. It records 7,520 emoticons, emojis, and words and their sentiment scores. **GI** (**Harvard General Inquirer)** is an early lexicon for text analysis: its basic spreadsheet has 11,788 entry words and their attributes, including positive, negative, strong, weak, active, passive, and so on [66]. The positive and negative attributes of all words are used for our study. **NRC** Emotion Lexicon [55] includes 14,182 words, their associated sentiments (binary), and emotion labels (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). The **OPL** (**Opinion Lexicon**) is created by [28], which includes 2,006 positive and 4,783 negative words. Finally, the **MPQA** (**Multi-Perspective Question Answerin**g) Subjectivity Lexicon developed by [69] has 8,222 words along with their POS tagging, polarity (positive, negative, or neutral), and intensity (strong or weak).

3.3 Knowledge-Enabled Transformer Models

We categorize the methods to integrate knowledge into three types: anterior, parallel, and posterior integration.

3.3.1 Anterior Integration. Anterior knowledge integration is the most popular approach, which is to study how to incorporate the knowledge into the sentence, such as forming a sentence tree with the branch being the incorporated knowledge and feeding it into transformer models like BERT. Anterior knowledge integration augments sentences with richer sentiment information and can be helpful for model training and fine-tuning. The main challenge of anterior integration is the potential risk of changing the meaning of the original sentence, and thus both [39] and [82] have introduced the soft-position and visible matrix to limit the impact of knowledge. We adopted the techniques introduced by [39] and [82] and studied the anterior integration of lexicon knowledge into financial sentiment analysis as a baseline. First, the sentence tree is constructed as shown in Figure 3, where w_i represents tokens in a sentence and w_{i1} and w_{i2} represent the queried lexical knowledge for w_i, which could be "is positive," "is neutral," or "is negative," for instance.

Figure 4 has illustrated how the embedding representation is generated from a sentence tree. The soft-position index is represented by the red number and the hard-position index is signified by the green number in the sentence tree. The token embedding is formed by flattening the tokens in the sentence tree into a sequence of tokens by their hard-position index. The position embedding is generated from the soft-position index along with the token embedding. The tokens in the original sentence are tagged as A, while the tokens in the auxiliary sentence are tagged as B for segment embedding.

3.3.2 Parallel Integration. Parallel Integration aims to develop a different model architecture for the knowledge base and train in parallel with pre-trained language models. Our parallel integration model architecture is illustrated in Figure 5. 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 the large language models (BERT/XLNet/RoBERTa) is a sentence pair. We use the same notations as by [67], which are $S = \{x_1, x_2, \ldots, x_l\}$ for a financial news headline or post sentence, and $aux(S)$ for its auxiliary sentence containing the corresponding targets and aspects. The auxiliary sentence takes a format of "*what do you think of aspect for target?*" or "*what do you think of target?*" Then, the input is in a format of "[cls] S [sep] $aux(S)$ [sep]" for BERT, and "<s> S </s> </s> $aux(S)$ </s>" for 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.[6] The scores are normalized to $[-1, +1]$ via min-max scaling, where -1

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

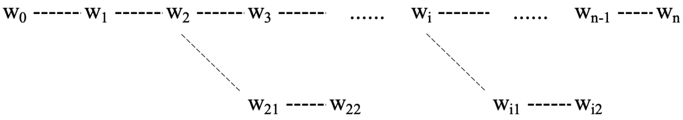
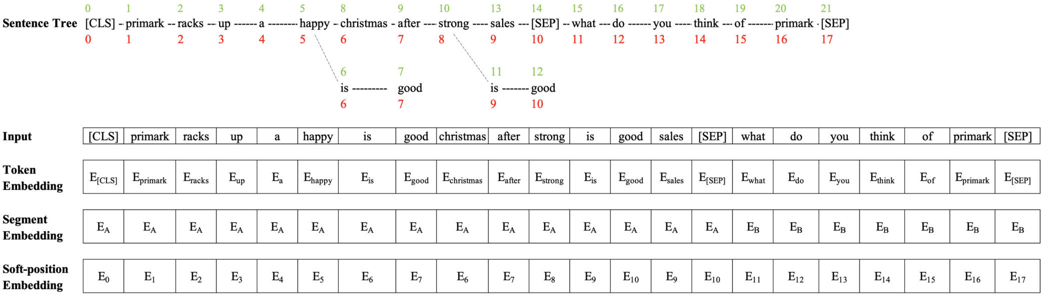

W0 ------ W1 ------ W2 ------ W3 ------ ...... Wi ------- ...... Wn-1 ----- Wn
W21 ----- W22
Wi1 ------ Wi2

Fig. 3. Construction of sentence tree.


Sentence Tree
0 1 2 3 4 5 8 9 10 13 14 15 16 17 18 19 20 21
[CLS] - primark -- racks -- up ------ a -------- happy - christmas - after --- strong ---- sales --- [SEP] -- what --- do ------ you ---- think --- of ------- primark · [SEP]
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
6 7
is --------- good
6 7
11 12
is ------- good
9 10

Input	[CLS]	primark	racks	up	a	happy	is	good	christmas	after	strong	is	good	sales	[SEP]	what	do	you	think	of	primark	[SEP]
Token Embedding	$E_{[CLS]}$	$E_{primark}$	E_{racks}	E_{up}	E_{a}	E_{happy}	E_{is}	E_{good}	$E_{christmas}$	E_{after}	E_{strong}	E_{is}	E_{good}	E_{sales}	$E_{[SEP]}$	E_{what}	E_{do}	E_{you}	E_{think}	E_{of}	$E_{primark}$	$E_{[SEP]}$
Segment Embedding	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B	E_B
Soft-position Embedding	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_6	E_7	E_8	E_9	E_{10}	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}

Fig. 4. The process of converting a sentence tree into an embedding representation for BERT.

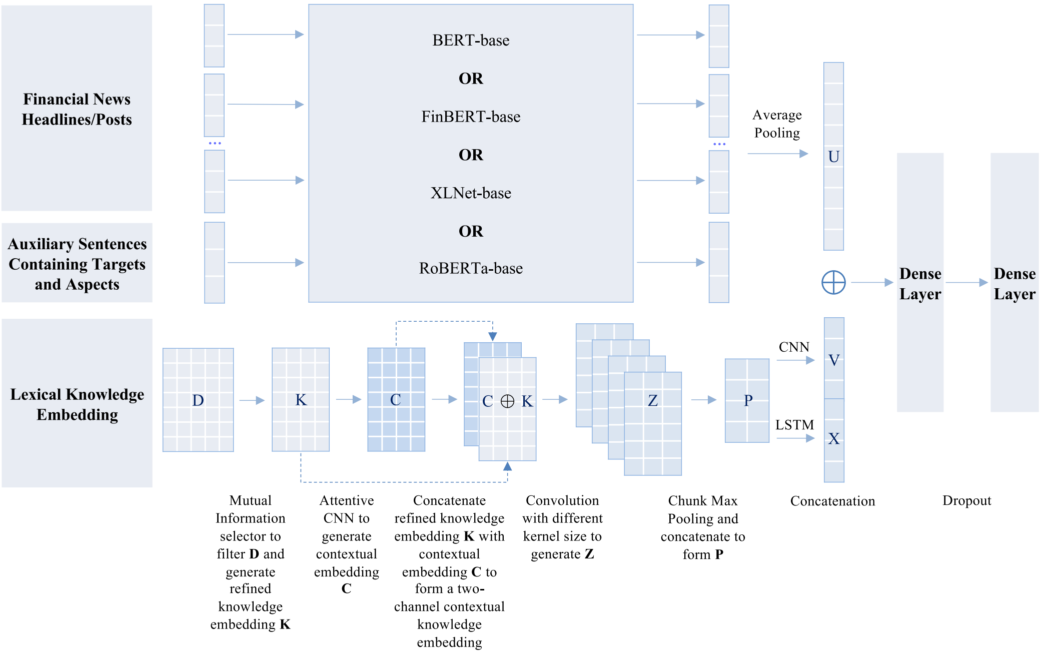

Financial News Headlines/Posts
Auxiliary Sentences Containing Targets and Aspects
Lexical Knowledge Embedding
BERT-base
OR
FinBERT-base
OR
XLNet-base
OR
RoBERTa-base
Average Pooling
U
⊕
Dense Layer
Dense Layer
CNN
LSTM
V
X
D
K
C
C ⊕ K
Z
P
Mutual Information selector to filter D and generate refined knowledge embedding K
Attentive CNN to generate contextual embedding C
Concatenate refined knowledge embedding K with contextual embedding C to form a two-channel contextual knowledge embedding
Convolution with different kernel size to generate Z
Chunk Max Pooling and concatenate to form P
Concatenation
Dropout

Fig. 5. Architecture of the proposed knowledge-enabled transformer models.

and +1 represent the most extreme sentiments. As an example, the scores for "happy" and "strong" are shown in Figure 6. The word "happy" has not only sentiment but also emotion score, in contrast to "strong" that usually carries only sentiment score. 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, returned with zeros. The coverage of master dictionary by lexicon on the SemEval 2017 Task 5 and FiQA Task 1 datasets are summarized in Figure 7.

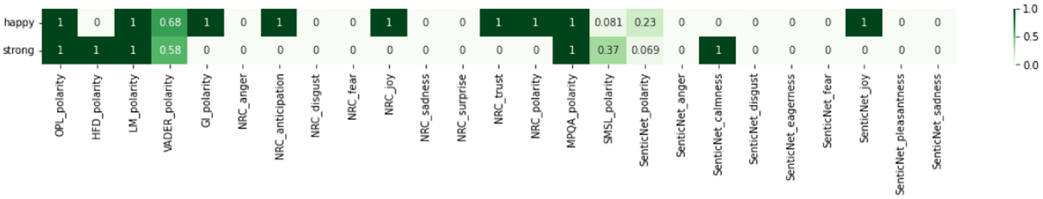
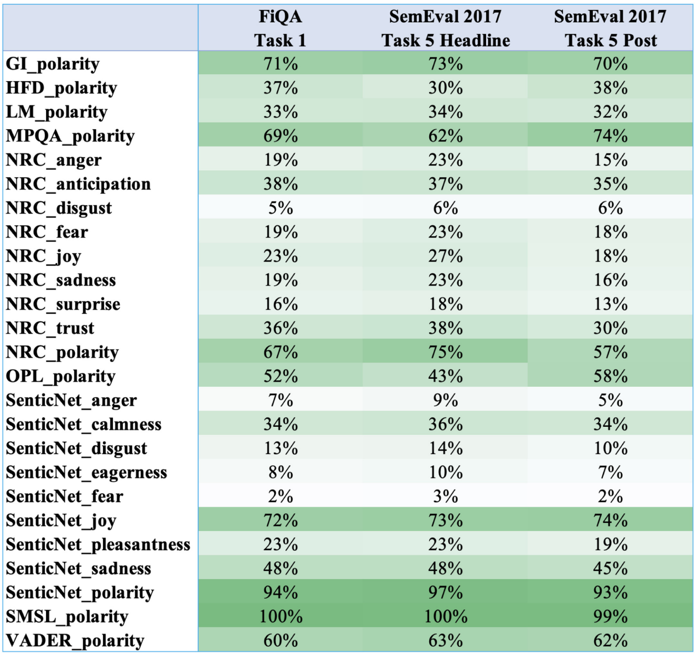


Fig. 6. Examples from master dictionary.



	FiQA Task 1	SemEval 2017 Task 5 Headline	SemEval 2017 Task 5 Post
GI_polarity	71%	73%	70%
HFD_polarity	37%	30%	38%
LM_polarity	33%	34%	32%
MPQA_polarity	69%	62%	74%
NRC_anger	19%	23%	15%
NRC_anticipation	38%	37%	35%
NRC_disgust	5%	6%	6%
NRC_fear	19%	23%	18%
NRC_joy	23%	27%	18%
NRC_sadness	19%	23%	16%
NRC_surprise	16%	18%	13%
NRC_trust	36%	38%	30%
NRC_polarity	67%	75%	57%
OPL_polarity	52%	43%	58%
SenticNet_anger	7%	9%	5%
SenticNet_calmness	34%	36%	34%
SenticNet_disgust	13%	14%	10%
SenticNet_eagerness	8%	10%	7%
SenticNet_fear	2%	3%	2%
SenticNet_joy	72%	73%	74%
SenticNet_pleasantness	23%	23%	19%
SenticNet_sadness	48%	48%	45%
SenticNet_polarity	94%	97%	93%
SMSL_polarity	100%	100%	99%
VADER_polarity	60%	63%	62%

Fig. 7. Statistics of coverage by master dictionary. % refers to the percentage of records in the dataset that have sentiment or emotion score from that lexicon.

Because the original knowledge embedding uses lexicons individually developed, it contains conflicting information and noise from alien language styles. To this end, we further apply feature selection techniques to 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 [17] or mutual information [6, 64]. Random forest measures the mean decrease in impurity, while mutual information captures various forms of dependency between variables, which is different from F-test, which captures linear dependency only. Mutual information is non-negative and a larger value represents higher dependency between the variables. We estimate mutual information based on entropy estimates from k-nearest neighbor distances ($k = 3$), as a larger k could introduce bias [36]. Mutual information is chosen in our study as [18] shows that the mutual information criterion can select features that minimize MSE and MAE in regressions. As illustrated in Figure 5, the mutual information selection is performed on lexical knowledge embedding $\boldsymbol{D}$ during the training process. Each text in the training dataset can be represented by a 25-dimensional embedding and also has a corresponding ground truth sentiment score. Our method involves computing the mutual information between ground truth sentiment scores and each of the 25 dimensions during training, ranking the importance of the dimensions, and selecting only lexical scores that have higher mutual information with ground truth sentiment score in the training dataset. For instance, the original lexical knowledge embedding of 25 dimensions will be reduced to 5 dimensions if the number of lexicons to be selected is set to 5. During this process, a refined knowledge embedding $\boldsymbol{K}$ will be generated, which will then be fed into an attentive CNN.

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 selected lexicons[7]:

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

Inspired by the literature on implementing the attention mechanism in deep neural networks, 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 emphasis, and thus each sentence also has a contextual embedding $\boldsymbol{C}(S) \in \mathbb{R}^{m \times n}$.

$$c_i = \sum_{i \neq j} \alpha_{i,j} \cdot x_j. \tag{2}$$

The attention weight $\alpha_{i,j}$ can be obtained by normalizing the score of a word pair $s(w_i, w_j)$ from a **Multi-Layer Perceptron** (**MLP**) through softmax function, where given $k = |i - j| - 1$ and $\lambda \in [0, 1)$, a decay factor to penalize the output score for reducing the impact of noise information that would be produced when the length of the sentence grows:

$$s(w_i, w_j) = (1 - \lambda)^k \cdot v_a^T tanh(W_a[x_i \oplus x_j]). \tag{3}$$

After that, we experimented with two methods to perform the subsequent convolution.

One-Dimensional Convolution. For one-dimensional convolution, the knowledge embedding is concatenated with contextual embedding directly to form a one-channel embedding and feed into CNNs to generate feature representation. The convolving kernel sizes are set to $d = 2, 3, 4, 5$, and the number of filters c is experimented with 4. Each convolution involves a filter $\boldsymbol{w} \in \mathbb{R}^{2n \times h}$, where n is the total number of sentiment and emotion scores across lexicons and h is the number of words for a sliding window. A new feature z_j, where $j \in [1, 2, \ldots, m - h + 1]$, is generated from a sliding window over words $x_{(j:j+h-1)}$ as

$$z_j = w_j \cdot K_{x_{(j:j+h-1)}} + b_j, \tag{4}$$

where b_{ij} is a bias term.

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

$$\boldsymbol{Z} = [\, z_{11} \cdots z_{1(m-h+1)} \,]. \tag{5}$$

The convolved features $\boldsymbol{Z}$ are activated by the ReLU function and chunk-max-pooled. 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 the pooled vector and $i \in [1, 2, 3, 4]$.

Two-Dimensional Convolution. The difference between one-dimensional and two-dimensional convolution is illustrated in Figure 8. Unlike one-dimensional convolution, which only captures global similarities and differences of lexicons, the two-dimensional convolution also can capture local characteristics of the most effective lexicons that are adjacent to each other. For two-dimensional convolution, we concatenate the knowledge embedding with contextual embedding to form a two-channel embedding and feed it into CNNs to generate feature representation. The

[7] In our case, the optimal n ranges from 3 to 8 out of the 25 dimensions.

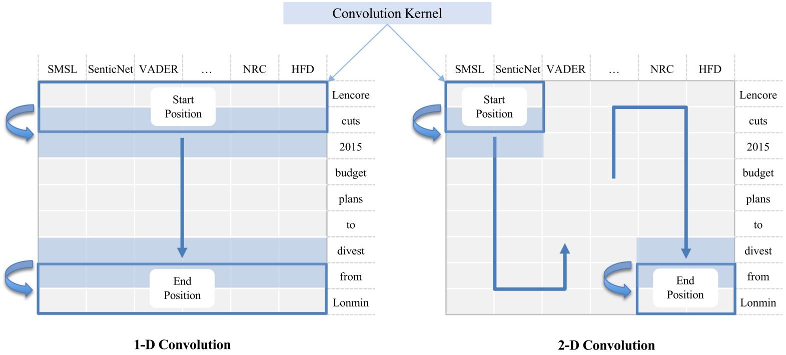


Fig. 8. One-dimensional and two-dimensional convolution.

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 $i \in [1,2,\ldots,n-e+1]$ and $j \in [1,2,\ldots,m-h+1]$, is generated from a sliding window over words $x_{(i:i+e-1,j:j+h-1)}$ as

$$z_{ij} = w_{ij} \cdot K_{x_{(i:i+e-1,\,j:j+h-1)}} + b_{ij}, \tag{6}$$

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

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1(m-h+1)} \\ z_{21} & \cdots & z_{2(m-h+1)} \\ \vdots & \ddots & \vdots \\ z_{(n-e+1)1} & \cdots & z_{(n-e+1)(m-h+1)} \end{bmatrix}. \tag{7}$$

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 the pooled vector and $i \in [1,2,3,4]$.

Subsequently, 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 [35] by using attentive convolution and chunk max pooling followed by additional CNN and LSTM layers in parallel to extract features further. 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

$$O = w_2 \cdot \sigma[w_1 \cdot \tanh(U \oplus V \oplus X) + b_1] + b_2, \tag{8}$$

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

To enable a fair comparison, all other settings are kept the same as the vanilla transformer models, except for the knowledge integration component.

3.3.3 *Posterior Integration.*

Posterior integration is defined as the addition of knowledge to the output embedding from transformer models. The most straightforward approach is a direct concatenation without further processing, which is formulated as follows and used in our study as a baseline:

$$O = w_2 \cdot \sigma[w_1 \cdot \tanh(U \oplus K) + b_1] + b_2, \tag{9}$$

where $\boldsymbol{U}$ is the output from the transformer model, $\boldsymbol{K}$ is the refined lexical knowledge embedding, and 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

A critical 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 [13, 14]. To deal with the catastrophic forgetting problem, Howard and Ruder [27] proposed three training techniques: slanted triangular learning rates, gradual unfreezing, and discriminative fine-tuning. A more recent method [10], however, recalls knowledge from pre-training without the original data by using a pre-training simulation mechanism and learns downstream tasks gradually by using an objective shifting mechanism. Specifically, it applies the idea of multi-task learning, which trains the model on the source and target tasks simultaneously to improve model performance with loss functions defined as follows:

$$Loss_M = \lambda Loss_T + (1 - \lambda) Loss_S, \tag{10}$$

where $Loss_T$ is the loss function for the target task, $Loss_S$ is the loss function for the source task, and $\lambda \in (0, 1)$ is a hyperparameter balancing the two tasks. $Loss_S$ optimizes the negative log posterior probability of the model parameters θ, given data of source tasks D_S. Similarly, $Loss_T$ optimizes the negative log posterior probability of the model parameters, given data of target tasks D_T.

The first challenge for multi-task learning is that the pre-training data is unavailable, and thus pre-training simulation is introduced as a quadratic penalty between the model parameters and the pre-trained parameters to approximate the optimization objective of the source task:

$$Loss_S = -logp(\theta | D_S) \approx \frac{1}{2} \gamma \sum\nolimits_i (\theta_i - \theta_i^*)^2, \tag{11}$$

where $\frac{1}{2}\gamma$ is the coefficient of quadratic penalty, θ is the model parameters, and θ^* is the local minimum of the parameter space of source task D_S.

The next challenge is that the optimization objective of adaptation is $Loss_T$, which is inconsistent with multi-task learning. To address it, the objective shifting is introduced to allow the loss function to shift to $Loss_T$ gradually with an annealing coefficient:

$$\begin{gathered} Loss_M = \lambda(t) Loss_T + (1 - \lambda(t)) Loss_S, \\ \lambda(t) = \frac{1}{1 + \exp(-k \cdot (t - t_0))}, \end{gathered} \tag{12}$$

where $\lambda(t)$ is computed as the sigmoid annealing function with $k \in (0, 1]$ and t_0 being the parameters controlling the annealing rate and timesteps, respectively.

This method has achieved state-of-the-art results on the benchmark datasets. Therefore, we apply this "recall-and-learn" training strategy [10] to prevent catastrophic forgetting in our fine-turning process for all pre-trained language models.

4 EXPERIMENTS

4.1 Datasets

The SemEval 2017 Task 5 dataset was developed for fine-grained sentiment analysis on financial news and microblogs [11]. The training data includes 1,142 financial news headlines and 1,694 posts with their target entities and corresponding sentiment scores but without aspects labeled.

Table 1. Level 1 and Level 2 Aspects in FiQA 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

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. The data were manually annotated by three independent financial experts according to the annotation guidelines defined by [11]. The final dataset was created by the fourth domain expert by consolidating the ratings. Inter-annotator agreements are used to assess the quality of the annotations. Specifically, for each pair of annotators, the Spearman's Rank Correlation on sentiment scores was computed, and then averaged across them [11]. 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 was from an open challenge [47], which consists of 498 financial news headlines and 675 posts with manually annotated target entities, aspects, and corresponding sentiment scores. 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. The following box is an example from the FiQA Task 1 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 [67, 70, 77] are chosen because many are developed for the (T)ABSA task and some achieved state-of-the-art results. Moreover, FinBERT [2] performed further pre-training to address the domain-specific language style and was ranked the first for sentiment analysis on Financial PhraseBank.[8] Similarly, the pre-trained language model input is a sentence pair, in which one sentence is the auxiliary sentence containing the target and aspect, and the other is the financial news headline or post. The average pooling is

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

performed on the last hidden state, followed by dropout. Finally, a last linear layer is added with the size of (768 × 1). The 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 a 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 the gold standard is not released, we perform a 10-fold cross-validation on two differently seeded runs for evaluation, and the mean score is reported. As for the 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 a maximum of 100 epochs, an initial learning rate of 3e-5 with a linear schedule with warm-up strategy, and Recall Adam as the optimizer.

5 RESULTS AND ANALYSIS

Consistent with previous studies [11, 47], we respectively report cosine similarities for SemEval 2017 Task 5, and MSE for FiQA Task 1 (see Tables 2 and 3). Additionally, we include R^2, which measures the percentage of variance explained by the model under evaluation. Under all columns and metrics, RoBERTa and XLNet outperform BERT and by significant margins even before the integration of sentiment knowledge. This confirms RoBERTa and XLNet as more effective deep representation models than BERT for the TABFSA task. Knowledge-enabled RoBERTa achieves state-of-the-art results on both SemEval 2017 Task 5 (Cosine[h] = 0.8495, 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 [1] (Cosine[h] = 0.7860, Cosine[p] = 0.7970), and circa 30% improvement from the previous best results produced by FinBERT on FiQA Task 1 by [2] (MSE = 0.07, R^2 = 0.55). From this perspective, the incorporation of financial knowledge can produce better model performance than retraining of finance domain-specific language models in TABFSA task. The results also show that overall parallel integration is more effective than posterior, which outperforms anterior integration. Notably, the anterior incorporation of multiple lexicon knowledge has decreased the model performance. In terms of 1D and 2D convolution in the parallel approach, they have produced comparable results, which means 1D convolution is more efficient because it requires less learnable parameters and training time.

5.1 Ablation Analysis

Ablation analysis is performed to validate the external knowledge embedding module. The results of models trained with 10 different random seeds for various transformer models and knowledge-enabled transformer models are provided in Tables 4 and 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. The knowledge selection through mutual information has further improved the model performance. Specifically, the FiQA Task 1 data has reported a 4% improvement in MSE for BERT with a 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. We have also included the CNN approach proposed by [35] and presented the results under k-RoBERTa (parallel-CNN w/MI [35]). Overall, the proposed k-RoBERTa (parallel-2D w/MI) still produces better or comparable results. Lastly, we have conducted a paired *t*-test between RoBERTa and

Table 2. Performance of Proposed Knowledge-Enabled Transformer Models in Comparison to the State-of-the-Art Approaches on SemEval 2017 Task 5

Model	**Headline**		**Post**	
	Cosine	R^2	Cosine	R^2
Lexicon-based [42]	0.1861	0.033	0.3032	0.052
Regression ensemble [33]	0.7100	-	0.7780	-
MLP ensemble [1]	0.7860	-	0.7970	-
FinBERT^a [2]	0.7969	0.635	0.7817	0.570
FinBERT^b [79]	0.7798	0.609	0.7626	0.536
BERT	0.7935	0.630	0.7886	0.581
k-BERT (anterior) [39, 82]	0.7809	0.610	0.7614	0.535
k-BERT (posterior)	0.7958	0.633	0.7903	0.584
k-BERT (parallel-1D)	0.7971	0.636	0.7916	0.587
k-BERT (parallel-2D)	0.7969	0.635	0.7912	0.586
XLNet	0.8199	0.676	0.8031	0.608
k-XLNet (anterior) [39, 82]	0.8014	0.644	0.7754	0.560
k-XLNet (posterior)	0.8215	0.675	0.8075	0.616
k-XLNet (parallel-1D)	0.8249	0.681	0.8074	0.615
k-XLNet (parallel-2D)	0.8270	0.685	0.8074	0.615
RoBERTa	0.8430	0.710	0.8085	0.617
k-RoBERTa (anterior) [39, 82]	0.8140	0.664	0.7754	0.560
k-RoBERTa (posterior)	0.8380	0.703	0.8063	0.614
k-RoBERTa (parallel-1D)	**0.8495**	**0.722**	**0.8113**	**0.623**
k-RoBERTa (parallel-2D)	**0.8483**	**0.721**	**0.8126**	**0.624**

Boldface indicates the top 2 result. We transcribe the results reported in [33] and [1]. "-" means not reported.

k-RoBERTa (parallel-2D w/MI), and the result shows that it is very significant for FiQA Task 1 (p = 0.002), significant for SemEval 2017 Task 5 Headline (p = 0.074), and marginally significant for SemEval 2017 Task 5 Post (p = 0.172).

5.2 Visualization of Attention

We visualize the average-pooled contextual embedding C generated by k-RoBERTa (parallel-2D) in Figure 9. A darker color means more attention is placed on the word. For example, the words "cut" and "divest" have been given more attention, which generally signifies a negative sentiment in finance. On the other hand, words such as "approve" and "lead" are typically positive sentiments in finance, which are also given more attention in our examples. Meanwhile, we also show the visualization of attention scores $s(w_i, w_j)$ produced by k-RoBERTa (parallel-2D) in Figure 10. Similarly, each row of the matrix represents a vector α, and a darker green cell indicates that more attention is being paid to the word in the corresponding column. As illustrated in Figure 10, the negative sentiment patterns abandon, cut, and divest are quite significant in the respective sentence. It can be concluded from Figure 10 that the correlation of a pair of words $s(w_i, w_j)$ can be understood as the degree to which w_i depends on w_j to indicate the sentiment of the corresponding sentence [83].

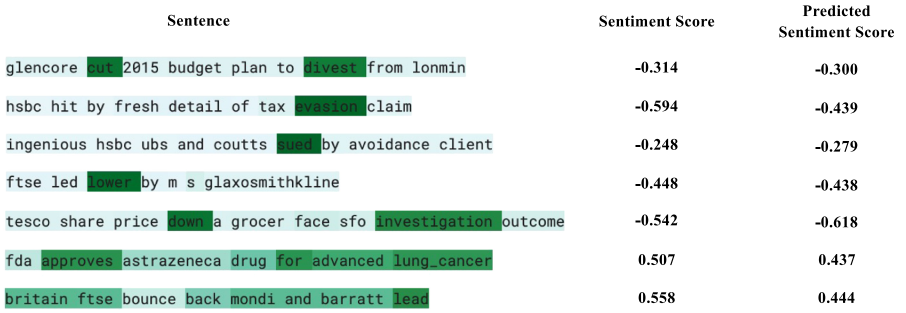
5.3 Knowledge Quality Analysis

One challenge in knowledge integration is called knowledge noise issue [39], which means too much knowledge integration may divert the sentence from the correct meaning. The precision

Table 3. Performance of Proposed Knowledge-Enabled Transformer Models in Comparison to State-of-the-Art Approaches in Sentiment Analysis Task on FiQA Task 1

Model	MSE	R^2
Lexicon-based [42]	0.1720	0.040
DNN ensemble [60]	0.0926	0.414
ULMFiT fine-tuning [78]	0.0800	0.400
FinBERT^a [2]	0.0700	0.550
FinBERT^b [79]	0.0636	0.613
BERT	0.0651	0.601
k-BERT (anterior) [39, 82]	0.0738	0.549
k-BERT (posterior)	0.0634	0.610
k-BERT (parallel-1D)	0.0624	0.616
k-BERT (parallel-2D)	0.0628	0.615
XLNet	0.0549	0.665
k-XLNet (anterior) [39, 82]	0.0627	0.619
k-XLNet (posterior)	0.0522	0.693
k-XLNet (parallel-1D)	0.0538	0.669
k-XLNet (parallel-2D)	0.0532	0.674
RoBERTa	0.0548	0.677
k-RoBERTa (anterior) [39, 82]	0.0602	0.642
k-RoBERTa (posterior)	0.0546	0.668
k-RoBERTa (parallel-1D)	**0.0499**	**0.705**
k-RoBERTa (parallel-2D)	**0.0490**	**0.711**

Boldface indicates the top 2 result. We transcribe the results reported in [60], [78], and [2].



Sentence	Sentiment Score	Predicted Sentiment Score
glencore cut 2015 budget plan to divest from lonmin	-0.314	-0.300
hsbc hit by fresh detail of tax evasion claim	-0.594	-0.439
ingenious hsbc ubs and coutts sued by avoidance client	-0.248	-0.279
ftse led lower by m s glaxosmithkline	-0.448	-0.438
tesco share price down a grocer face sfo investigation outcome	-0.542	-0.618
fda approves astrazeneca drug for advanced lung_cancer	0.507	0.437
britain ftse bounce back mondi and barratt lead	0.558	0.444

Fig. 9. Visualization of average-pooled contextual embedding C generated by k-RoBERTa (parallel-2D).

and coverage of lexicons impact the effectiveness of external knowledge integration into the fine-tuning process. It is observed that anterior integration is more sensitive to noise knowledge. In contrast, for parallel and posterior integration, with the increase in the number of lexicon scores incorporated by mutual information, the model performance 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 complement

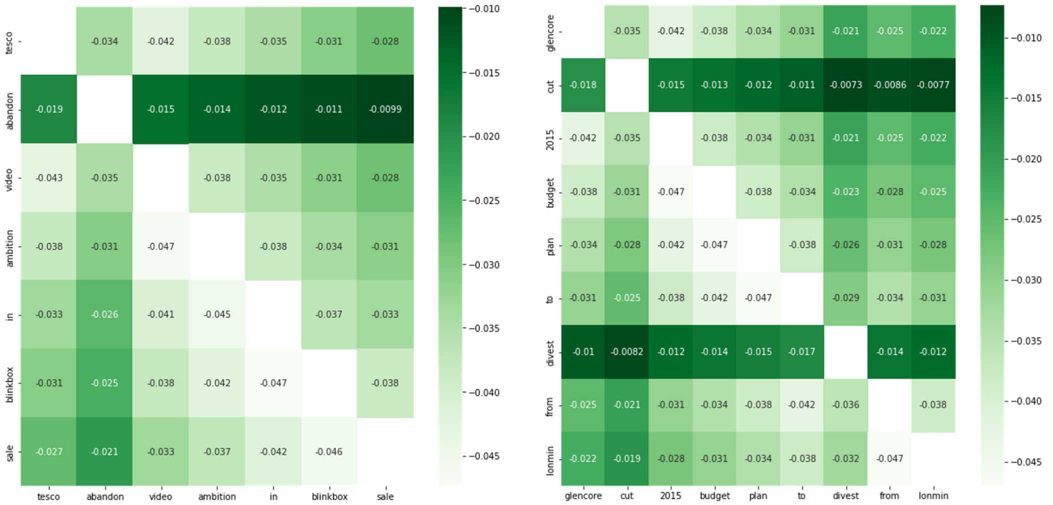


Fig. 10. Visualization of attention scores $s(w_i, w_j)$ generated by k-RoBERTa (parallel-2D).

Table 4. Ablation Analysis for SemEval 2017 Task 5

Cosine	**Headline**			**Post**		
Similarity	Mean	Median	SD	Mean	Median	SD
BERT	0.7935	0.7904	0.0096	0.7886	0.7850	0.0108
k-BERT (parallel-2D w/o MI)	0.7932	0.7932	**0.0064**	0.7889	0.7888	0.0118
k-BERT (parallel-2D w/ MI)	**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 (parallel-2D w/o MI)	0.7954	0.7977	0.0063	0.7822	0.7813	**0.0086**
k-FinBERT (parallel-2D w/ MI)	**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 (parallel-2D w/o MI)	0.8261	0.8260	0.0083	0.8067	0.8066	0.0108
k-XLNet (parallel-2D w/ MI)	**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 (parallel-2D w/o MI)	0.8462	0.8462	**0.0048**	0.8117	0.8116	0.0175
k-RoBERTa (parallel-CNN w/ MI [35])	0.8455	0.8481	0.0090	**0.8128**	**0.8122**	**0.0110**
k-RoBERTa (parallel-2D w/ MI)	**0.8483**	**0.8500**	0.0170	0.8126	0.8118	0.0125

w/ MI means Mutual Information is adopted to select lexicons. w/o MI means all lexicons are used without any selection. The parallel-2D is our proposed model and parallel-CNN means the CNN proposed by [35] is adopted.

the learning process. Moreover, the closer to the accuracy bound of the deep neural network, the more challenging 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 Figure 11). In terms of selected lexicon scores, the experiment shows that sentiment and emotion knowledge are helpful, though generally sentiment scores are more critical than emotion scores. Furthermore, the importance of finance domain-specific lexicons such as LM and SMSL are consistently higher than most of the general-purpose lexicons. In particular, SMSL, HFD, and LM sentiment contribute to the best model performance in FiQA Task 1 and SMSL, SenticNet, and VADER sentiment contribute to the best model performance in the SemEval 2017 Task 5 Post dataset. The model performance is

Table 5. Ablation Analysis for FiQA Task 1 Sentiment Analysis

MSE	Mean	Median	SD
BERT	0.0651	0.0602	0.0191
k-BERT (parallel-2D w/o MI)	0.0647	0.0628	**0.0168**
k-BERT (parallel-2D w/ MI)	**0.0628**	**0.0573**	0.0180
FinBERT	0.0675	0.0668	0.0172
k-FinBERT (parallel-2D w/o MI)	0.0672	0.0679	0.0163
k-FinBERT (parallel-2D w/ MI)	**0.0646**	**0.0623**	**0.0157**
XLNet	0.0549	0.0526	0.0147
k-XLNet (parallel-2D w/o MI)	0.0544	0.0528	0.0143
k-XLNet (parallel-2D w/ MI)	**0.0532**	**0.0502**	**0.0119**
RoBERTa	0.0548	0.0526	**0.0173**
k-RoBERTa (parallel-2D w/o MI)	0.0511	0.0488	0.0176
k-RoBERTa (parallel-CNN w/ MI [35])	0.0500	0.0447	0.0185
k-RoBERTa (parallel-2D w/ MI)	**0.0490**	**0.0420**	0.0185

w/ MI means Mutual Information is adopted to select lexicons. w/o MI means all lexicons are used without any selection. The parallel-2D is our proposed model and parallel-CNN means the CNN proposed by [35] is adopted.

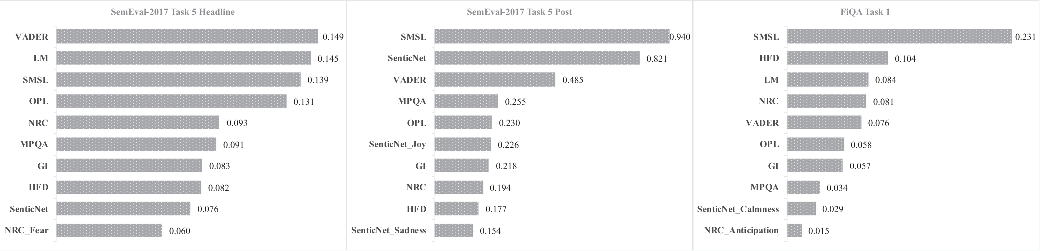


Fig. 11. Mutual information of lexicons for SemEval 2017 Task 5 and FiQA Task 1 datasets.

decreased after other lexicons are added subsequently. Meanwhile, the SemEval 2017 Task 5 Headline dataset has the best model performance when VADER, LM, SMSL, OPL, NRC, and MPQA sentiment are integrated. It is observed that the general-purpose lexicon also plays a critical role such as VADER in the SemEval-2017 Task 5 Headline and the SenticNet in the SemEval-2017 Task 5 Post dataset. As for emotion dimensions, joy, sadness, and fear tend to be more relevant for the TABFSA task.

5.4 Case Study

In most cases, incorporating external knowledge is beneficial for the accuracy of predicted sentiment scores. However, 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. Essentially, the lexicons selected by mutual information are not only relevant at a general level but also highly correlated with this particular sentence.

In Scenario 2, however, knowledge-enabled RoBERTa is no better than the standalone RoBERTa. For this concrete example, although the snippet of "invalidated by US court" is negative, the word "invalidated" does not carry any sentiment in two out of the three 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 the words "drug" and "court" in this context is considered noise knowledge, mentioned earlier. Due to the existence of noise knowledge, lexicons that are selected by mutual information are more relevant at a broad level, but less correlated with this specific sentence.

```
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 in parallel. Incorporating external knowledge into transformer models has achieved state-of-the-art performance on the SemEval 2017 Task 5 and the FiQA Task 1 datasets. Meanwhile, we have discovered and demonstrated that parallel integration is a more effective approach than anterior and posterior when multiple sources of lexical knowledge are incorporated. Lastly, the results show that incorporating financial and general lexicon knowledge can improve model performances more than retraining finance domain-specific language models in the TABFSA task. We plan to investigate three further issues in future work: (1) the 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

[1] 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*. 540–546.

[2] Dogu Araci. 2019. FinBert: Financial sentiment analysis with pre-trained language models. https://arxiv.org/abs/1908.10063.

[3] 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*. 124–128.

[4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

[5] 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*. 253–259.

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

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[8] 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 *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20)*. 105–114.

[9] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC'22)*. 3829–3839.

[10] 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 *Proceedings of EMNLP'20*. 7870–7881.

[11] 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*.

[12] 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*. 1967–1971.

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

[14] 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. https://arxiv.org/abs/2002.06305.

[15] Zhendong Dong and Qiang Dong. 2003. HowNet-a hybrid language and knowledge resource. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 820–824.

[16] Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving Earth one tweet at a time through the lens of artificial intelligence. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN'22)*. 1–9. https://doi.org/10.1109/IJCNN55064.2022.9892271

[17] Minhaz Bin Farukee, M. S. Zaman Shabit, Md Rakibul Haque, and A. H. M. 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*. 118–134.

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

[19] Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of AAAI*. 10681–10689.

[20] 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*. 899–903.

[21] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab@ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-Generated Text*. 146–153.

[22] 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 (2017), 139–152.

[23] Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*. 94–104.

[24] Kai He, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Meta-based self-training and re-weighting for aspect-based sentiment analysis. *IEEE Transactions on Affective Computing* (2023). https://doi.org/10.1109/TAFFC.2022.3202831

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

[26] 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 (2019), 1174–1209.

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

[28] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 168–177.

[29] 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 *Proceedings of EMNLP'20*. 6989–6999.

[30] Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM'14*.

[31] 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*. 1961–1966.

[32] Fuwei Jiang, Joshua Lee, Xiumin Martin, and Guofu Zhou. 2019. Manager sentiment and stock returns. *Journal of Financial Economics* 132, 1 (2019), 126–149.

[33] 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*. 888–893.

[34] Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2017. RiTUAL-UH at SemEval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-'17)*. Association for Computational Linguistics, 877–882. https://doi.org/10.18653/v1/S17-2150

[35] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1746–1751.

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

[37] Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems* 235 (2022), 107643.

[38] Bing Liu. 2015. *Sentiment Analysis—Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

[39] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908.

[40] 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. https://arxiv.org/abs/1907.11692.

[41] 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 *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'20)*. 4513–4519.

[42] Steven Loria, P. Keen, M. Honnibal, and R. Yankovsky. 2021. TextBlob: Simplified Text Processing—TextBlob 0.16. 0 documentation. *TextBlob: Simplified Text Processing*. Available online: https://textblob.readthedocs.io/en/dev/. Accessed August 6, 2021.

[43] 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 (2011), 35–65.

[44] 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 *Proceedings of IJCAI'18*. 4244–4250.

[45] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. 2023. Multi-source aggregated classification for stock price movement prediction. *Information Fusion* 91 (2023), 515–528.

[46] Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Proceedings of AAAI'18*. 5876–5883.

[47] 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 Web Conference 2018*. 1941–1942.

[48] 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 (2014), 782–796.

[49] Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini, and Jacopo Staiano. 2017. Fortia-FBK at SemEval-2017 task 5: Bullish or Bearish? Inferring sentiment towards brands from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, 817–822. https://doi.org/10.18653/v1/S17-2138

[50] Rui Mao and Xiao Li. 2021. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13534–13542.

[51] Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. MetaPro: A computational metaphor processing model for text pre-processing. *Information Fusion* 86-87 (2022), 30–43. https://doi.org/10.1016/j.inffus.2022.06.002

[52] Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 1222–1231.

[53] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing* (2023). https://doi.org/10.1109/TAFFC.2022.3204972

[54] 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*. 3111–3119.

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

[56] 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 (2013), 621–633.

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

[58] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

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

[60] 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*. 1973–1977.

[61] Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING'16*. 1546–1556.

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

[63] 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*. Association for Computational Linguistics.

[64] 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 (2018), 1–25.

[65] Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: A lexicon for emotion analysis from crowd-annotated news. https://arxiv.org/abs/1405.1605.

[66] 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.

[67] Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT'19*. 380–385.

[68] 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 (2015), 4999–5010.

[69] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP'05*. 347–354.

[70] Zhengxuan Wu and Desmond C. Ong. 2021. Context-guided BERT for targeted aspect-based sentiment analysis. In *Proceedings of AAAI'21*. 14094–14102.

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

[72] Frank Xing, Erik Cambria, and Yue Zhang. 2019. Sentiment-aware volatility forecasting. *Knowledge Based Systems* 176 (2019), 68–76.

[73] Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proceedings of COLING'20*. 978–987.

[74] Frank Z. Xing, Erik Cambria, and Roy E. Welsch. 2018. Intelligent asset allocation via market sentiment views. *IEEE Computational Intelligence Magazine* 13, 4 (2018), 25–34.

[75] Frank Z. Xing, Filippo Pallucchini, and Erik Cambria. 2019. Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management* 56, 3 (2019), 554–564.

[76] Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. CN-DBpedia: A never-ending Chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 428–438.

[77] 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'19*. 2324–2335.

[78] Steve Yang, Jason Rosenfeld, and Jacques Makutonin. 2018. Financial aspect-based sentiment analysis using deep representations. https://arxiv.org/abs/1808.07931.

[79] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. FinBERT: A pretrained language model for financial communications. https://arxiv.org/abs/2006.08097.

[80] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS'19*. 5754–5764.

[81] 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 (2015), 531–549.

[82] Anping Zhao and Yu Yu. 2021. Knowledge-enabled BERT for aspect-based sentiment analysis. *Knowledge-Based Systems* (2021), 107220.

[83] Zhiwei Zhao and Youzheng Wu. 2016. Attention-based convolutional neural networks for sentence classification. In *Proceedings of INTERSPEECH'16*. 705–709.

Received 24 March 2022; revised 19 December 2022; accepted 26 December 2022